CENGAGE Learning

INTRODUÇÃO À ECONOMETRIA

UMA ABORDAGEM MODERNA

TRADUÇÃO DA 4ª EDIÇÃO NORTE-AMERICANA

JEFFREY M. WOOLDRIDGE

Outras Obras

Economia Internacional
Robert J. Garbaugh

Estatística Aplicada à Administração e Economia
Anderson, Sweeney e Williams

Introdução à Economia – Tradução da 3ª edição norte-americana
N. Gregory Mankiw

Macroeconomia – Princípios e Aplicações
Robert E. Hall e Marc Lieberman

Microeconomia – Princípios e Aplicações
Robert E. Hall e Marc Lieberman

Matemática Aplicada à Administração, Economia e Contabilidade
Afranio Murolo e Giácomo Bonetto

Princípios de Economia – 5ª edição revista
Carlos Roberto Martins Passos e Otto Nogami

INTRODUÇÃO À ECONOMETRIA

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Wooldridge, Jeffrey M., 1960-
Introdução à econometria : uma abordagem moderna / Jeffrey M. Wooldridge ; tradução Rogério Cézar de Souza, José Antônio Ferreira ; revisão técnica Nelson Carvalheiro. -- São Paulo : Thomson Learning, 2007.

Título original: Introductory econometrics : a modern approach
1. reimpr. da 1. ed. de 2006
Bibliografia
ISBN 85-221-0421-2

1. Econometria I. Carvalheiro, Nelson. II. Título.

05-8002 CDD-330.015195

Índices para catálogo sistemático:

1. Econometria 330.015195

INTRODUÇÃO À ECONOMETRIA

Uma Abordagem Moderna

Jeffrey M. Wooldridge

Michigan State University

Tradução

Rogério César de Souza
José Antônio Ferreira

Revisão Técnica

Nelson Carvalheiro
Doutor em Economia pela USP e Professor Titular
do Departamento de Economia da FEA/PUC-SP

THOMSON

Austrália Brasil Canadá Cingapura Espanha Estados Unidos México Reino Unido

THOMSON

Gerente Editorial:
Dulcy Grisolia

Editora de Desenvolvimento:
Tatiana Pavanelli Valsi

Supervisora de Produção Editorial:
Patricia La Rosa

Produtor Editorial:
Fábio Gonçalves

Produtora Gráfica:
Fabiana Alencar Albuquerque

Título original:
Introductory Econometrics: A Modern Approach – 2th edition
(ISBN: 0-324-11364-1)

Tradutores:
Rogério César de Souza (Capítulos 1-5),
José Antônio Ferreira

Revisão Técnica:
Nelson Carvalheiro

Copidesque:
Peterso Roberto Rissatti

Revisão:
Andréa Medeiros
Ana Paula Ribeiro,
Silvana Gouveia

Editoração Eletrônica:
ERJ – Composição Editorial e Artes Gráficas Ltda.

Capa:
Eduardo Bertolini

Impresso no Brasil.
Printed in Brazil.
1 2 3 4 09 08 07

Condomínio E-Business Park
Rua Werner Siemens, 111
Prédio 20 – Espaço 03
Lapa de Baixo
CEP 05069-900
São Paulo – SP
Tel.: (11) 3665-9900
Fax: (11) 3665-9901
sac@thomsonlearning.com.br
www.thomsonlearning.com.br

Dados Internacionais de Catalogação na Publicação (CIP) (Câmara Brasileira do Livro, SP, Brasil)
Wooldridge, Jeffrey M., 1960-
Introdução à econometria : uma abordagem moderna / Jeffrey M. Wooldridge ; tradução Rogério Cézar de Souza, José Antônio Ferreira ; revisão técnica Nelson Carvalheiro. — São Paulo : Thomson Learning, 2007.
Título original: Introductory econometrics : a modern approach
1. reimpr. da 1. ed. de 2006
Bibliografia
ISBN 85-221-0421-2
1. Econometria I. Carvalheiro, Nelson.
II. Título.
05-8002
CDD–330.015195
Índices para catálogo sistemático:
1. Econometria 330.015195

Prefácio

Minha motivação para escrever *Introdução à Econometria — Uma Abordagem Moderna* vem de uma lacuna, razoavelmente ampla, que identifiquei entre como a econometria é ensinada nos cursos de graduação e o que os pesquisadores empíricos pensam sobre os métodos econométricos e suas aplicações. Com igual importância, convenci-me de que ensinar econometria introdutória da perspectiva dos usuários profissionais da econometria simplificaria, de fato, a apresentação, além de tornar o assunto mais interessante.

Baseado em numerosas reações positivas à primeira edição norte-americana, parece que minha filosofia sobre como ensinar a econometria introdutória é partilhada por muitas pessoas. É gratificante que professores de formações e interesses variados — incluindo a microeconomia aplicada, a macroeconomia aplicada, a análise de política econômica, a ciência política e a econometria teórica —, ensinando estudantes com níveis de preparação muito diferentes, tenham abraçado a abordagem moderna da econometria adotada neste livro. Conseqüentemente, a estrutura desta primeira edição brasileira é muito parecida com a da primeira edição norte-americana. A ênfase ainda está em aplicar a econometria aos problemas do mundo real. Todo método econométrico é motivado por uma questão particular com a qual o pesquisador se defronta ao analisar dados não-experimentais. O foco principal da obra está em entender e interpretar as hipóteses à luz das aplicações empíricas reais: a matemática requerida não vai além da álgebra dos cursos de graduação e da probabilidade e estatística básicas.

ORGANIZADO PARA O ECONOMETRISTA DE HOJE

A característica mais perceptível desta edição, que distingue este texto de outros, é a separação dos tópicos por tipo de dados analisados. Essa é uma diferença clara em relação à abordagem tradicional, que apresenta um modelo linear, lista todas as hipóteses que podem ser necessárias em algum ponto futuro da análise e, então, prova ou afirma resultados sem conectá-los claramente às hipóteses. Minha abordagem é, em primeiro lugar, tratar, na Parte Um, da análise de regressão múltipla com dados de corte transversal, sob a hipótese de amostragem aleatória. Essa estrutura é completamente natural para os estudantes, pois eles estão familiarizados com ela desde os cursos de estatística introdutória. O mais importante é que ela nos permite distinguir hipóteses feitas sobre o modelo de regressão da população subjacente — hipóteses que podem ter um conteúdo econômico determinado ou um conteúdo comportamental geral — de hipóteses sobre como os dados foram extraídos para compor uma amostra.

As discussões sobre as conseqüências da amostragem não-aleatória podem ser tratadas de um modo intuitivo, após os estudantes terem um bom domínio do modelo de regressão múltipla aplicado a amostras aleatórias.

Uma característica importante de uma abordagem moderna é que as variáveis explicativas — com a variável dependente — são tratadas como resultados de variáveis aleatórias. Para as ciências sociais, admitir variáveis explicativas aleatórias é muito mais realista do que a hipótese tradicional de variáveis explicativas não-aleatórias. Um benefício importante é que a abordagem modelo populacional/amostragem aleatória que utilizo reduz bastante o número de hipóteses que os estudantes devem absorver e entender. Ironicamente, a abordagem clássica da análise de regressão, que trata as variáveis explicativas como valores fixos em amostras repetidas e está difundida nos livros introdutórios, aplica-se literalmente a dados coletados em uma estrutura experimental. Além disso, as contorções exigidas para formular e explicar as hipóteses podem ser confusas para os estudantes.

Meu foco sobre o modelo populacional enfatiza que as hipóteses fundamentais subjacentes à análise de regressão, tal como a hipótese de média zero dos fatores não-observados, estão apropriadamente formuladas, condicionadas às variáveis explicativas. Isso leva a um entendimento claro dos tipos de problemas, tal como a heteroscedasticidade (variância não-constante), que podem invalidar os procedimentos padrões da inferência. Adicionalmente, pude eliminar várias interpretações erradas que surgem nos textos de econometria em todos os níveis. Apenas para citar alguns exemplos, explico a razão de o R-quadrado usual ainda ser válido como uma medida do grau de ajuste na presença de heteroscedasticidade (e, mais adiante, na presença dos erros serialmente correlacionados nas equações de séries de tempo); discuto, em um nível bastante intuitivo, porque os testes para a forma funcional não devem ser vistos como testes gerais de variáveis omitidas não observadas; e posso facilmente explicar porque sempre se deve incluir, em um modelo de regressão, variáveis extras de controle que são não-correlacionadas com a variável explicativa de interesse (tal como uma variável de decisão).

Como as hipóteses da análise de corte transversal são relativamente diretas e realistas, os estudantes ficam envolvidos mais cedo com aplicações sérias de corte transversal, sem ter de se preocupar com as questões espinhosas de tendência, sazonalidade, correlação serial, alta persistência e regressão espúria que aparecem em abundância nos modelos de regressão de séries temporais. Inicialmente, imaginei que meu tratamento da regressão com dados de corte transversal, seguida pela regressão com dados de séries de tempo, cairia nas boas graças dos professores cujos interesses de pesquisa estão na microeconomia aplicada, e parece que esse é o caso. Tem sido gratificante que aqueles que adotaram este livro e que têm uma inclinação para as séries de tempo ficaram igualmente entusiasmados com a estrutura da obra. Ao postergar a análise econométrica de séries temporais, pude colocar o foco apropriado sobre as armadilhas potenciais da análise de dados de séries de tempo que não surgem com dados de corte transversal. Com efeito, a econometria de séries temporais obteve, por fim, o tratamento sério que ela merece em um livro introdutório.

Para esta edição, escolhi conscientemente os tópicos que são importantes para ler artigos de revistas e para realizar pesquisas empíricas básicas. Dentro de cada tópico, omiti deliberadamente muitos testes e procedimentos de estimação que, embora tradicionalmente incluídos nos livros-texto, não têm resistido ao teste empírico do tempo. Da mesma forma, enfatizei os tópicos mais recentes que têm se mostrado claramente úteis, tal como obter estatísticas de teste que são robustas em relação à heteroscedasticidade (ou à correlação serial) cuja forma é desconhecida, usar dados de vários anos para a análise de decisão, ou resolver o problema de variáveis omitidas pelo método de variáveis instrumentais. Parece que fiz as escolhas corretas, pois me lembro de bem poucas sugestões para acrescentar ou excluir material, especialmente nos capítulos da regressão básica das Partes Um e Dois. Um tópico que, de fato, expandi foi a estimação dos mínimos desvios absolutos (MDA) do Capítulo 9. O MDA está se tornando cada vez mais popular nos trabalhos empíricos, especialmente quando a distribuição condicional da variável dependente é assimétrica ou tem caudas largas. Ao ler pesquisas empíricas da área da economia do trabalho, da economia do setor público e outros campos, é provável que os estudantes encontrem cada vez mais modelos estimados por MDA.

. Ao reescrever alguns trechos desta edição, tentei aperfeiçoar mais a abordagem sistemática. O termo sistemática significa que cada tópico está fundamentado, de um modo lógico, no material anterior, e as hipóteses são introduzidas somente se forem necessárias para obter uma conclusão. Por exemplo, os pesquisadores aplicados, bem como os teóricos, sabem que nem todas as hipóteses de Gauss-Markov são necessárias para mostrar que os estimadores de mínimos quadrados ordinários (MQO) são não-viesados. Contudo, quase todos os livros de econometria introduzem o conjunto completo de hipóteses (muitas das quais são redundantes ou, em alguns casos, logicamente conflitantes) antes de provar a inexistência de viés do MQO. De forma semelhante, a hipótese de normalidade é incluída entre as hipóteses que são necessárias para o Teorema de Gauss-Markov, quando é razoavelmente bem conhecido que a normalidade não desempenha nenhum papel para mostrar que os estimadores de MQO são os melhores estimadores lineares não-viesados.

Minha abordagem sistemática estende-se para o estudo das propriedades de amostras grandes, em que as hipóteses de consistência são introduzidas somente quando necessárias. Isso torna relativamente fácil cobrir tópicos mais avançados, como usar cortes transversais agrupados, explorar as estruturas de dados de painel e aplicar os métodos de variáveis instrumentais. Trabalhei para fornecer uma visão unificada da econometria, de acordo com a qual todos os estimadores e as estatísticas de testes são obtidas usando um pouco de princípios, intuitivamente racionais, de estimação e testes (os quais, evidentemente, também têm justificativas rigorosas). Por exemplo, os testes básicos de regressão para a heteroscedasticidade e a correlação serial são fáceis de ser compreendidos pelos estudantes porque eles já têm um conhecimento sólido de regressão. Isso contrasta com os tratamentos que fornecem um conjunto de receitas desconexas para procedimentos econométricos ultrapassados.

Como, ao longo deste livro, enfatizo as relações *ceteris paribus*, vou diretamente para a análise de regressão múltipla após abordar somente um capítulo do modelo de regressão simples. Isso motiva os estudantes a pensar sobre aplicações sérias mais cedo. Também atribuo muito mais destaque à análise de decisão utilizando todos os tipos de estruturas. Os tópicos práticos, como usar variáveis *proxy* para obter efeitos *ceteris paribus* e obter erros-padrão dos efeitos parciais nos modelos com termos de interação, são discutidos de modo simples.

NOVO NESTA EDIÇÃO

No Capítulo 3, há uma discussão completa do viés de variáveis omitidas no modelo de regressão múltipla; acaba não sendo especialmente difícil caracterizar o viés no caso geral. O apêndice do Capítulo 3 contém uma derivação do viés que requer somente álgebra e estatística básicas.

O Capítulo 6 contém uma discussão mais detalhada de modelos com termos de interação. Percebi que os estudantes podem ter dificuldades ao interpretar os parâmetros de tais modelos, de modo que tentei dar melhores explicações. Os Capítulos 7 e 13 apresentam mais detalhes sobre como computar os testes de Chow para os modelos de regressão entre grupos de diferentes unidades de corte transversal e entre diferentes períodos de tempo. Nos Capítulos 8 e 12, forneço um argumento explícito e simples sobre porque o R-quadrado ainda é válido como uma medida do grau de ajuste quando o modelo sofre de heteroscedasticidade ou de correlação serial. O Capítulo 9 inclui uma discussão simples de amostragem estratificada básica, um tópico que nasceu de um curso de segundo semestre no qual ministrei aulas.

O Capítulo 17 contém material complexo, e expandi a discussão sobre como interpretar modelos não-lineares de variável dependente limitada, inclusive como eles podem ser comparados com as estimativas de modelos lineares padrão. Novos gráficos foram usados para ilustrar as comparações.

O livro traz mais de 80 bancos de dados disponíveis para *download* pela senha 4212 na página deste livro no site da Thomson (www.thomsonlearning.com.br), dos quais muitos permitem comparar

as estimativas dos retornos de freqüentar cursos profissionalizantes e cursos de graduação; testar se a poupança dos planos de pensão norte-americanos substituem outras formas de poupança ou representa poupança nova; testar se os restaurantes *fast-food* praticam discriminação de preços contra minorias; testar se o casamento afeta a produtividade ou a remuneração dos jogadores profissionais de basquetebol; estudar o efeito de mais escolhas sobre os investimentos em planos de pensão; estimar os efeitos dos gastos públicos em escolas sobre o desempenho dos estudantes; testar se as leis de uso do cinto de segurança e de limite de velocidade afetam as taxas de acidentes e de mortes; e estimar funções de demanda para maçãs e peixe ecologicamente produzidos. Além disso, os exercícios para computador — disponíveis no site da Thomson — foram expandidos para explorar os novos bancos de dados. Alguns dos conjuntos de dados não são utilizados no livro. Em vez disso, eles podem ser usados em problemas, em exames ou para servir de base para um trabalho de final de curso.

PROJETADO PARA ESTUDANTES DE CURSOS DE GRADUAÇÃO E ADAPTADO PARA ESTUDANTES DE CURSOS DE PÓS-GRADUAÇÃO

O livro é direcionado para estudantes de cursos de graduação em economia que estudaram álgebra e um semestre de probabilidade e estatística introdutórias. (Os Apêndices A, B, C, D, E e F — disponíveis no site da Thomson — contêm o material de pré-requisito.) Não se espera que um curso de econometria de um semestre ou de um trimestre abranja tudo, ou mesmo alguma parte, dos tópicos mais avançados da Parte Três. Um curso de introdução típico incluiria os Capítulos 1 a 8, os quais compreendem as bases das regressões simples e múltipla para dados de corte transversal. Dado que a ênfase está na intuição e na interpretação dos exemplos empíricos, o material dos oito primeiros capítulos deveria ser colocado à disposição dos estudantes dos cursos de graduação na maioria dos departamentos de economia. A maioria dos professores também vai querer cobrir pelo menos partes dos capítulos sobre análise de regressão com dados de séries de tempo (Capítulos 10, 11 e 12), com graus variados de profundidade. No curso de um semestre em que leciono, no Estado de Michigan, trabalho o Capítulo 10 cuidadosamente, dou uma visão geral do material do Capítulo 11 e abordo o material sobre correlação serial do Capítulo 12. Acredito que esse curso básico de um semestre fornece ao aluno fundamentos sólidos para escrever trabalhos empíricos, como uma monografia de final de curso ou um texto para um seminário. O Capítulo 9 contém tópicos mais especializados que surgem ao analisar dados de corte transversal, incluindo problemas de dados tais como *outliers* e amostragem não aleatória. Para um curso de um semestre, esse capítulo pode ser abandonado sem perda de continuidade.

A estrutura do livro é ideal para um curso com foco em corte transversal/análise de decisão: os capítulos de séries de tempo podem ser abandonados, dando lugar a tópicos dos Capítulos 9, 13, 14 ou 15. O Capítulo 13 é "avançado" somente no sentido de que ele trata de duas novas estruturas de dados: cortes transversais independentemente agrupados e análise de dados de painel para dois períodos. Tais estruturas de dados são especialmente úteis para análise de decisão, e esse capítulo fornece vários exemplos a esse respeito. Os estudantes com um bom domínio dos Capítulos 1 a 8 terão pouca dificuldade com o Capítulo 13. O Capítulo 14 trata dos métodos de dados de painel mais avançados e provavelmente será coberto somente em um segundo curso. Uma boa maneira de finalizar um curso sobre métodos de corte transversal é compreender os rudimentos da estimação de variáveis instrumentais do Capítulo 15.

Tenho utilizado material selecionado da Parte Três, incluindo os Capítulos 13, 15 e 17, nos seminários do curso de graduação direcionados para produzir um trabalho de pesquisa sério. Além do curso básico de um semestre, os estudantes que foram expostos à análise básica de dados de painel, de estimação de variáveis instrumentais e de modelos de variável dependente limitada estão preparados para

ler boa parte da literatura aplicada das ciências sociais. O Capítulo 17 apresenta uma introdução aos modelos mais comuns de variável dependente limitada.

O livro também é adequado para um curso introdutório de pós-graduação, no qual a ênfase está mais nas aplicações do que nas derivações que usam álgebra matricial. Além disso, para os professores que querem apresentar a matéria na forma matricial, os apêndices D e E abordam, de modo autosuficiente, a álgebra matricial e o modelo de regressão múltipla na forma matricial. O Apêndice E abrange a análise assimptótica em profundidade maior para estudantes avançados.

No Estado de Michigan, os estudantes dos cursos de doutorado das muitas áreas que requerem análise de dados — incluindo contabilidade, economia agrícola, economia do desenvolvimento, finanças, economia internacional, economia do trabalho, macroeconomia, ciência política e finanças públicas — descobriram que o livro é uma ponte útil entre o trabalho empírico que eles lêem e a econometria mais teórica que eles aprendem no nível de doutoramento.

CARACTERÍSTICAS BÁSICAS

Os professores e estudantes parecem apreciar as questões formuladas no texto, cujas respostas encontram-se no Apêndice F. Essas questões têm a intenção de dar ao estudante um retorno imediato sobre seu desempenho. Cada capítulo contém muitos exemplos. Vários deles são estudos de caso retirados de artigos publicados recentemente, levemente modificados para simplificar a análise, sem sacrificar seus principais pontos.

Os problemas de final de capítulo e exercícios para computador — disponíveis no site da Thomson — são totalmente orientados para o trabalho empírico, em vez das derivações complicadas. Os estudantes são solicitados a fundamentar cuidadosamente suas respostas, com base no que aprenderam. Os exercícios para computador expandem, em geral, os exemplos do texto. Vários exercícios usam bancos de dados de trabalhos publicados ou conjuntos de dados similares que são motivados por pesquisas publicadas em economia e em outros campos.

Uma característica única deste livro é o extensivo glossário. As definições e descrições curtas serão um lembrete de grande auxílio para os alunos que estejam estudando para exames ou lendo pesquisas empíricas que usam métodos econométricos.

CONJUNTOS DE DADOS

Mais de 80 bancos de dados estão disponíveis em Excel e em ".DES", arquivos de texto que descrevem as variáveis e que podem ser abertos no programa Excel. A maioria dos conjuntos de dados é proveniente de pesquisas reais, de modo que alguns são bastante grandes. Exceto quando for o caso ilustrar as várias estruturas de dados, os bancos de dados não são descritos no texto. Este livro está direcionado para um curso em que o trabalho com o computador desempenha papel importante. Os conjuntos de dados podem ser encontrados na página deste livro no site da Thomson.

Um banco de dados em Access com os nomes e as informações dos dados estatísticos também está disponível no site e permite ao leitor uma busca por nome do arquivo. Os nomes de diversas variáveis foram traduzidos e adaptados ao longo do texto e estão grafados com acentos (como *salário*, por exemplo) para facilitar a memorização do leitor brasileiro. Entretanto, como os dados estatísticos em Excel estão em inglês, a busca pelo banco de dados permite ao leitor encontrar o nome da variável em inglês e sua busca no respectivo arquivo em Excel.

SUGESTÕES PARA MONTAR SEU CURSO

Já comentei sobre o conteúdo da maioria dos capítulos e possíveis estruturas de cursos. Aqui, farei alguns comentários mais específicos sobre o material dos capítulos que podem ser abordados ou postergados.

O Capítulo 9 tem exemplos interessantes (tal como uma regressão que inclui a pontuação do QI como uma variável explicativa). Os nomes das variáveis *proxy* não devem ser formalmente apresentados para descrever esses tipos de exemplos, e costumo apresentá-los quando termino a análise de corte transversal. No Capítulo 12, para um curso de um semestre, não apresento o material sobre inferência robusta na presença de correlação serial quando estou tratando da análise de mínimos quadrados ordinários, bem como de modelos dinâmicos de heteroscedasticidade.

Mesmo em um segundo curso, prefiro despender pouco tempo no Capítulo 16, que trata da análise de equações simultâneas. Se há uma questão sobre a qual as pessoas divergem é a importância das equações simultâneas. Alguns consideram que esse material é fundamental; outros pensam que é raramente aplicável. Minha visão é que os modelos de equações simultâneas são demasiadamente utilizados (veja o Capítulo 16 para uma discussão). Se lermos os trabalhos aplicados cuidadosamente, variáveis omitidas e erros de medida são provavelmente uma das maiores razões para adotar a estimação de variáveis instrumentais, e é por isso que uso variáveis omitidas para motivar a estimação de variáveis instrumentais no Capítulo 15. Além disso, os modelos de equações simultâneas são indispensáveis para estimar funções de demanda e oferta, e eles também são aplicáveis em alguns outros casos importantes.

O Capítulo 17 é o único que considera modelos inerentemente não-lineares em seus parâmetros, e isso impõe uma carga adicional para o estudante. O primeiro material que deveria ser tratado nesse capítulo são os modelos de resposta binária probit e logit. Minha apresentação dos modelos Tobit e de regressão censurada ainda parecem originais: reconheço explicitamente que o modelo Tobit é aplicável a resultados de solução de canto em amostras aleatórias, ao passo que a regressão censurada é aplicável quando o processo de coleta de dados censura a variável dependente.

O Capítulo 18 trata de alguns tópicos mais recentes da econometria de séries de tempo, inclusive o teste de raízes unitárias e a cointegração. Abordo esse material somente no segundo semestre de um curso, seja no nível de graduação ou de pós-graduação. Uma introdução razoavelmente detalhada para a previsão também está incluída no Capítulo 18.

O Capítulo 19, que deveria ser acrescentado ao programa de cursos que exigem um trabalho de conclusão, é muito mais extensivo que capítulos semelhantes de outros livros. Ele resume alguns métodos apropriados para vários tipos de problemas e estruturas de dados, aponta dificuldades potenciais, explica com algum detalhe como escrever um trabalho de conclusão de curso em economia empírica e inclui sugestões de possíveis projetos.

AGRADECIMENTOS

Gostaria de agradecer aqueles que revisaram a primeira ou a segunda edição norte-americana, fizeram extensos comentários ou influenciaram de alguma maneira o livro. São eles:

Richard Agnello
University of Delaware

Scott Baier
Clemson University

Eli Berman
Boston University

James Cardon
Brigham Young University

Rogério César de Souza
Universidade Paulista

Amitabh Chandra
Dartmouth University

Christopher Cornwell
University of Georgia

Edward Coulson
Pennsylvania State University

William Even
Miami University of Ohio

Adrian Fleissig
St. Louis University

Arthur Goldberger
University of Wisconsin

Daniel Hamermesh
University of Texas

Bruce Hansen
University of Wisconsin

KyungSo I in
Wichita State University

Datelina Ivanovo
SUNY, Albany

Heejoon Kang
Indiana University

Manfred Keil
Claremont McKenna College

Neha Khanna
SUNY, Binghamton

Esfandiar Maasoumi
Southern Methodist University

Kristin McCue
Texas A &M University

Philip Meguire
University of Canterbury

John Mullahy
University of Wisconsin

William Neils
Texas A &M University

David Neumark
Michigan State University

Leslie Papke
Michigan State University

Soo-Bin Park
Carleton University

Jeffrey Pliskin
Hamilton College

Joseph Quinn
Boston College

Nagesh Revankar
SUNY, Buffalo

Louise Russell
Rutgers University

Shinichi Sakata
University of Michigan

Mark Showalter
Brigham Young University

Jeffrey Smith
University of Maryland

John Spitzer
SUNY, Brockport

Leanna Stiefel
New York University

Wendy Stock
Montana State University

Christopher Tuber
Northwestern University

George Tavlas
Bank of Greece

Larry Taylor
Lehigh University

Pravin Trivedi
Indiana University

Robert Trost
George Washington University

Hiroki Tsurumi
Rutgers University

Robert Turner
Colgate University

Timothy Vogelsang
Cornell University

Melvyn Weeks
University of Cambridge

Diana Whistler
University of British Columbia

Paul Wilson
University of Texas

Keith Womar
University of Mississippi

Jeffrey Zabel
Tufts University

Considerei muitos dos seus comentários sobre a primeira edição norte-americana, mesmo que não tenha alterado o material de acordo com alguma preferência específica. Em alguns casos, as opiniões dos comentaristas e dos usuários conflitaram, e decidi deixar a organização do livro como está. Em outros casos, continuarei a cogitar sugestões específicas feitas por um ou mais comentaristas. Naturalmente, estou aberto a sugestões sobre possíveis melhorias futuras.

Muitos estudantes e assistentes de professores encontraram erros na primeira edição norte-americana ou sugeriram a reformulação de alguns parágrafos. O número de tais colaboradores é tão grande que não há espaço aqui para listá-los. Gostaria de agradecer os esforços de Chirok Han, que leu cuidadosamente as provas da primeira edição, e a Ali Becker, que verificou esta edição.

Uma vez mais, adorei trabalhar com o pessoal da South-Western/Thomson Learning. Coletivamente, eles conduziram a realização desta edição de modo gentil, mas com mão firme. Meu editor de desenvolvimento, Andy McGuire, ajudou-me a refinar muitos comentários de colaboradores e leitores sobre a primeira edição, e foi de grande auxílio ao construir uma estratégia de revisão do livro. Starratt Alexander encarregou-se da tarefa onerosa de editor de produção e fez um trabalho excepcional. Peggy Buskey e Vicky True cuidaram, com grande habilidade, do site do livro. Uma vez mais, Malvine Litten e seu grupo do LEAP, e em particular Rachel Morris, fizeram um trabalho magnífico de digitação e edição do manuscrito.

Esta edição ainda é dedicada à minha mulher, Leslie, e aos nossos filhos, Edmund e Gwenyth. Leslie fez comentários valiosos sobre o livro e identificou alguns erros de impressão. Além disso, ela continua me encorajando muito, apesar de ter ficado bastante cética quando eu disse, após a publicação da primeira edição norte-americana, o quanto estava feliz por ter finalizado o "projeto".

Sobre o Autor

Jeffrey M. Wooldridge é professor emérito de Economia na Universidade Estadual de Michigan, onde leciona desde 1991. De 1986 a 1991, foi professor-assistente de Economia no Massachusetts Institute of Technology. Obteve seu bacharelado, com especialização em Ciência da Computação e Economia, na Universidade da Califórnia, Berkeley, em 1982, e seu doutorado em Economia na Universidade da Califórnia, San Diego, em 1986. Publicou mais de duas dezenas de artigos em revistas internacionalmente reconhecidas e muitos capítulos de livros. É autor de *Econometric Analysis of Cross Section and Panel Data*. Seus prêmios incluem: um Alfred P. Sloan Research Fellowship, o Plura Scripsit da *Econometric Theory*; o Sir Richard Stone do *Journal of Applied Econometrics*; e três prêmios de professor do ano da pós-graduação do MIT. Além de ser membro do *Journal of Econometrics*, é também editor do *Journal of Business and Economic Statistics*, co-editor de econometria do *Economics Letters* e participa do corpo editorial do *Journal of Econometrics* e da *Review of Economics and Statistics*. Ocasionalmente, também atua como consultor de econometria para a Arthur Andersen, de Chicago, e para a Charles River Associates, de Boston.

Sumário

Capítulo 1 A Natureza da Econometria e dos Dados Econômicos 1

1.1 O que é Econometria? 1
1.2 Passos na Análise Econômica Empírica 2
1.3 A Estrutura dos Dados Econômicos 5
Dados de Corte Transversal 5
Dados de Séries de Tempo 8
Cortes Transversais Agrupados 9
Dados de Painel ou Longitudinais 10
Um Comentário sobre Estruturas de Dados 12
1.4 A Causalidade e a Noção de *Ceteris Paribus* na Análise Econométrica 12
Resumo 17

PARTE 1
ANÁLISE DE REGRESSÃO COM DADOS DE CORTE TRANSVERSAL 19

Capítulo 2 O Modelo de Regressão Simples 20

2.1 Definição do Modelo de Regressão Simples 20
2.2 Derivação das Estimativas de Mínimos Quadrados Ordinários 25
Uma Nota sobre Terminologia 34
2.3 Mecânica do Método MQO 34
Valores Estimados e Resíduos 34
Propriedades Algébricas das Estatísticas de MQO 36
Grau de ajuste 38
2.4 Unidades de Medida e Forma Funcional 39
Os Efeitos de Mudanças das Unidades de Medida sobre as Estatísticas de MQO 40
Incorporação de Não-Linearidades na Regressão Simples 41
O Significado da Regressão "Linear" 44
2.5 Valores Esperados e Variâncias dos Estimadores de MQO 45
Inexistência de Viés em MQO 45
Variâncias dos Estimadores de MQO 51
Estimação da Variância do Erro 55

2.6	Regressão através da Origem	58
	Resumo	59
	Problemas	59
	Apêndice 2A	62
Capítulo 3	**Análise de Regressão Múltipla: Estimação**	**64**
3.1	Funcionabilidade da Regressão Múltipla	64
	Modelo com duas Variáveis Independentes	64
	Modelo com k *Variáveis Independentes*	67
3.2	Mecânica e Interpretação dos Mínimos Quadrados Ordinários	69
	Obtenção das Estimativas de MQO	69
	Interpretação da Equação de Regressão de MQO	70
	Sobre o Significado de "Manter Outros Fatores Fixos" na Regressão Múltipla	73
	Variação de mais de uma Variável Independente Simultaneamente	73
	Valores Estimados e Resíduos de MQO	74
	Interpretação de "Parcialidade" da Regressão Múltipla	75
	Comparação das Estimativas das Regressões Simples e Múltipla	75
	Grau de Ajuste	77
	Regressão através da Origem	79
3.3	O Valor Esperado dos Estimadores de MQO	80
	Inclusão de Variáveis Irrelevantes em um Modelo de Regressão	85
	Viés de Variável Omitida: O Caso Simples	86
	Viés de Variável Omitida: Casos mais Gerais	90
3.4	A Variância dos Estimadores de MQO	91
	Os Componentes das Variâncias de MQO: Multicolinearidade	92
	Variâncias em Modelos Mal Especificados	96
	Estimação de σ^2*: Os Erros-Padrão dos Estimadores de MQO*	97
3.5	Eficiência de MQO: O Teorema de Gauss-Markov	99
	Resumo	100
	Problemas	101
	Apêndice 3A	105
Capítulo 4	**Análise de Regressão Múltipla: Inferência**	**110**
4.1	Distribuições Amostrais dos Estimadores de MQO	110
4.2	Testes de Hipóteses sobre um único Parâmetro Populacional: O Teste *t*	113
	Teste contra Hipóteses Alternativas Unilaterais	116
	Teste contra Hipóteses Alternativas Bilaterais	121
	Testes de outras Hipóteses sobre β_j	123
	Cálculos dos p*-Valores dos Testes* t	126
	Lembrete sobre a Linguagem do Teste de Hipóteses Clássico	129
	Significância Econômica ou Prática versus *Significância Estatística*	129
4.3	Intervalos de Confiança	131
4.4	Testes de Hipóteses sobre uma Combinação Linear dos Parâmetros	134
4.5	Testes de Restrições Lineares Múltiplas: O Teste *F*	137
	Teste de Restrições de Exclusão	137
	Relação entre as Estatísticas F *e* t	143

A Forma R-*quadrado da Estatística* F 144
Cálculo dos p-*Valores para Testes* F 146
A Estatística F *para a Significância Geral de uma Regressão* 147
Teste de Restrições Lineares Gerais 148
4.6 Descrição dos Resultados da Regressão 149
Resumo 152
Problemas 152

Capítulo 5 Análise de Regressão Múltipla: MQO Assimptótico 158

5.1 Consistência 158
A Derivação da Inconsistência no Método MQO 161
5.2 Normalidade Assimptótica e Inferência de Amostras Grandes 163
Outros Testes de Amostras Grandes:
A Estatística Multiplicador de Lagrange 166
5.3 Eficiência Assimptótica de MQO 169
Resumo 171
Problemas 171
Apêndice 5A 172

Capítulo 6 Análise de Regressão Múltipla: Problemas Adicionais 174

6.1 Efeitos da Dimensão dos Dados nas Estatísticas MQO 174
Os Coeficientes Beta 177
6.2 Um pouco mais sobre a Forma Funcional 179
Um pouco mais sobre o Uso de Formas Funcionais Logarítmicas 179
Modelos com Funções Quadráticas 182
Modelos com Termos de Interação 187
6.3 Um pouco mais sobre o Grau de Ajuste e a Seleção de Regressores 189
O R-*Quadrado Ajustado* 190
O Uso do R-*quadrado Ajustado para a Escolha entre*
Modelos Não-Aninhados 191
O Controle de muitos Fatores na Análise de Regressão 193
A Adição de Regressores para Reduzir a Variância do Erro 194
6.4 Previsão e Análise de Resíduos 195
Intervalos de Confiança de Previsões 195
Análise de Resíduos 199
Previsão de y *quando a Variável Dependente é log*(y) 201
Resumo 203
Problemas 204

Capítulo 7 Análise de Regressão Múltipla com Informações Qualitativas: Variáveis Binárias (ou *Dummy*) 207

7.1 A Descrição das Informações Qualitativas 207
7.2 Uma Única Variável *Dummy* Independente 209
A Interpretação dos Coeficientes de Variáveis Dummy Explicativas
quando a Variável Dependente é Expressa como log(y) 214
7.3 O Uso de Variáveis *Dummy* para Categorias Múltiplas 216
Incorporação de Informações Ordinais com o Uso de Variáveis Dummy 218

7.4 Interações Envolvendo Variáveis *Dummy* 221
Interações entre Variáveis Dummy 221
Consideração de Inclinações Diferentes 223
Verificação de Diferenças nas Funções de Regressão entre Grupos 227
7.5 Uma Variável Dependente Binária: O Modelo de Probabilidade Linear 230
7.6 Um Pouco mais sobre Análise e Avaliação de Políticas e Programas Governamentais 236
Resumo 238
Problemas 239

Capítulo 8 Heteroscedasticidade 243

8.1 Conseqüências da Heteroscedasticidade para o Método MQO 243
8.2 Inferência Robusta em Relação à Heteroscedasticidade após e Estimação MQO 244
Computando Testes LM *Robustos em Relação à Heteroscedasticidade* 249
8.3 O Teste da Existência de Heteroscedasticidade 251
O Teste de White para a Heteroscedasticidade 254
8.4 Estimação de Mínimos Quadrados Ponderados 256
A Heteroscedasticidade É Percebida como uma Constante Multiplicativa 256
A Necessidade de Estimar a Função de Heteroscedasticidade: O MQG Factível 262
8.5 O Modelo de Probabilidade Linear Revisitado 266
Resumo 269
Problemas 270

Capítulo 9 Problemas Adicionais de Especificação e de Dados 272

9.1 Má Especificação da Forma Funcional 272
O Teste RESET como um Teste Geral da Má Especificação da Forma Funcional 275
Testes Contra Alternativas Não-Aninhadas 277
9.2 Utilizando Variáveis Proxy para Variáveis Explicativas Não-Observadas 278
O Uso de Variáveis Dependentes Defasadas como Variáveis Proxy 283
9.3 Propriedades do Método MQO quando há Erros de Medida 285
Erro de Medida na Variável Dependente 285
Erro de Medida em uma Variável Explicativa 287
9.4 Ausência de Dados, Amostras Não-Aleatórias e Observações Extremas 292
Ausência de Dados 292
Amostras Não-Aleatórias 293
Observações Extremas ou Atípicas 295
Resumo 300
Problemas 301

PARTE 2
ANÁLISE DE REGRESSÃO COM DADOS DE SÉRIES TEMPORAIS 305

Capítulo 10 O Básico da Análise de Regressão com Dados de Séries Temporais 306

10.1 A Natureza dos Dados das Séries Temporais 306

10.2 Exemplos de Modelos de Regressão de Séries Temporais 307
Modelos Estáticos 308
Modelos de Defasagens Distributivas Finitas 308
Convenção sobre o Índice Temporal 311
10.3 Propriedades de Amostra Finita do MQO sob as Hipóteses Clássicas 311
Inexistência de Viés do MQO 311
As Variâncias dos Estimadores MQO e o Teorema de Gauss-Markov 315
Inferência sob as Hipóteses do Modelo Linear Clássico 318
10.4 Forma Funcional, Variáveis *Dummy* e Números-Índices 320
10.5 Tendência e Sazonalidade 327
Caracterização de Séries Temporais com Tendência 327
Uso de Variáveis com Tendência na Análise de Regressão 330
Interpretação sobre a Retirada da Tendência de Regressões com a Inclusão de uma Tendência Temporal 333
Cálculo do R-*Quadrado quando a Variável Dependente Apresenta Tendência* 334
Sazonalidade 336
Resumo 338
Problemas 338

Capítulo 11 Questões Adicionais quanto ao Uso do MQO com Dados de Séries Temporais 340

11.1 Séries Temporais Estacionárias e Fracamente Dependentes 340
Séries Temporais Estacionárias e Não-Estacionárias 341
Séries Temporais Fracamente Dependentes 342
11.2 Propriedades Assimptóticas do MQO 345
11.3 O Uso de Séries Temporais Altamente Persistentes na Análise de Regressão 352
Séries Temporais Altamente Persistentes 353
Transformações de Séries Temporais Altamente Persistentes 357
A Decisão sobre uma Série de Tempo Ser I(1) 358
11.4 Modelos Dinamicamente Completos e a Ausência de Correlação Serial 360
11.5 A Hipótese de Homoscedasticidade para Modelos de Séries Temporais 363
Resumo 364
Problemas 365

Capítulo 12 Correlação Serial e Heteroscedasticidade em Regressões de Séries Temporais 368

12.1 As Propriedades do MQO com Erros Serialmente Correlacionados 368
Inexistência de Viés e Consistência 368
Eficiência e Inferência 369
O Grau de Ajuste 370
A Correlação Serial na Presença da Variável Dependente Defasada 371
12.2 O Teste da Correlação Serial 372
O Teste t *de Correlação Serial AR(1) com Regressores Estritamente Exógenos* 373
O Teste de Durbin-Watson sob as Hipóteses Clássicas 375

O Teste da Correlação Serial AR(1) sem Regressores Estritamente Exógenos 376
O Teste da Correlação Serial de Ordem mais Elevada 378
12.3 A Correção da Correlação Serial com Regressores Estritamente Exógenos 380
A Obtenção do Melhor Estimador Linear Não-Viesado no Modelo AR(1) 380
A Estimação MQG Factível com Erros AR(1) 382
Comparação entre MQO e MQGF 384
A Correção da Correlação Serial para Ordens mais Elevadas 386
12.4 Diferenciação e Correlação Serial 387
12.5 Inferência Robusta em Relação à Correlação Serial após o MQO 388
12.6 Heteroscedasticidade em Regressões de Séries Temporais 392
Estatísticas Robustas em Relação à Heteroscedasticidade 392
O Teste da Heteroscedasticidade 393
A Heteroscedasticidade Condicional Auto-Regressiva 394
Heteroscedasticidade e Correlação Serial em Modelos de Regressão 396
Resumo 397
Problemas 398

PARTE 3
TÓPICOS AVANÇADOS **401**

Capítulo 13 O Agrupamento de Cortes Transversais ao Longo do Tempo. Métodos Simples de Dados de Painel **402**

13.1 O Agrupamento Independente de Cortes Transversais ao Longo do Tempo 403
O Teste de Chow de Mudança Estrutural ao Longo do Tempo 407
13.2 Análise de Decisões Governamentais com Agrupamentos de Cortes Transversais 408
13.3 Análise de Dados de Painel de dois Períodos 414
A Organização dos Dados de Painel 420
13.4 Análise de Decisões Governamentais com Dados de Painel de dois Períodos 421
13.5 A Diferenciação com mais de dois Períodos de Tempo 424
Resumo 429
Problemas 430
Apêndice 13A 431

Capítulo 14 Métodos Avançados de Dados de Painel **433**

14.1 Estimação de Efeitos Fixos 433
A Regressão das Variáveis Dummy 437
Efeitos Fixos ou Primeira Diferenciação? 439
Efeitos Fixos com Painéis Não Equilibrados 440
14.2 Modelos de Efeitos Aleatórios 441
Efeitos Aleatórios ou Efeitos Fixos? 445
14.3 A Aplicação de Métodos de Dados de Painel a outras Estruturas de Dados 445
Resumo 447
Problemas 448
Apêndice 14A 449

Capítulo 15 Estimação de Variáveis Instrumentais e Mínimos Quadrados de dois Estágios 453

15.1 Motivação: Variáveis Omitidas em um Modelo de Regressão Simples 454
Inferência Estatística com o Estimador de VI 457
Propriedades da VI com uma Variável Instrumental Pobre 462
O Cálculo do R-*Quadrado após a Estimação de VI* 464
15.2 Estimação de VI do Modelo de Regressão Múltipla 464
15.3 Mínimos Quadrados de dois Estágios 468
Uma Única Variável Explicativa Endógena 468
Multicolinearidade e MQ2E 471
Variáveis Explicativas Endógenas Múltiplas 472
O Teste de Hipóteses Múltiplas após a Estimação por MQ2E 473
15.4 Soluções de VI de Problemas de Erros nas Variáveis 473
15.5 O Teste de Endogeneidade e o Teste de Restrições Sobreidentificadoras 475
O Teste de Endogeneidade 475
O Teste de Restrições Sobreidentificadoras 477
15.6 O MQ2E com Heteroscedasticidade 478
15.7 A Aplicação do MQ2E a Equações de Séries Temporais 479
15.8 A Aplicação do MQ2E em Cortes Transversais Agrupados e em Dados de Painel 481
Resumo 484
Problemas 484
Apêndice 15A 488

Capítulo 16 Modelos de Equações Simultâneas 491

16.1 A Natureza dos Modelos de Equações Simultâneas 491
16.2 Viés de Simultaneidade no MQO 496
16.3 A Identificação e a Estimação de uma Equação Estrutural 498
A Identificação em um Sistema de Duas Equações 498
Estimação por MQ2E 503
16.4 Sistemas com mais de duas Equações 505
Identificação em Sistemas com três ou mais Equações 505
Estimação 506
16.5 Modelos de Equações Simultâneas com Séries Temporais 506
16.6 Modelos de Equações Simultâneas com Dados de Painel 510
Resumo 513
Problemas 514

Capítulo 17 Modelos com Variáveis Dependentes Limitadas e Correções da Seleção Amostral 517

17.1 Modelos Logit e Probit de Resposta Binária 518
A Especificação de Modelos Logit e Probit 518
Estimação de Máxima Verossimilhança de Modelos Logit e Probit 521
Testes de Hipóteses Múltiplas 522
A Interpretação das Estimativas Logit e Probit 523
17.2 O Modelo Tobit para Resposta de Solução de Canto 529

A Interpretação das Estimativas Tobit 531
Problemas de Especificação nos Modelos Tobit 536
17.3 O Modelo de Regressão de Poisson 537
17.4 Modelos de Regressão Censurada e Truncada 542
Modelos de Regressão Censurada 543
Modelos de Regressão Truncada 547
17.5 Correções da Seleção Amostral 549
Quando o MQO é Consistente na Amostra Selecionada? 549
Truncamento Ocasional 551
Resumo 555
Problemas 556
Apêndice 17A 558

Capítulo 18 Tópicos Avançados sobre Séries Temporais 559

18.1 Modelos de Defasagem Distribuída Infinita 560
A Defasagem Distribuída Geométrica (ou de Koyck) 562
Modelos de Defasagem Distribuída Racional 564
18.2 O Teste de Raízes Unitárias 567
18.3 Regressão Espúria 572
18.4 Co-Integração e Modelos de Correção de Erro 574
Co-Integração 574
Modelos de Correção de Erro 579
18.5 Previsão 581
Tipos de Modelos de Regressão Utilizados na Previsão 583
Previsão um Passo à Frente 584
A Comparação de Previsões um Passo à Frente 588
Previsão com Múltiplos Passos à Frente 589
A Previsão de Processos com Tendência, Sazonais e Integrados 592
Resumo 597
Problemas 599

Capítulo 19 A Montagem de um Projeto na Prática 602

19.1 A Formulação de uma Pergunta 602
19.2 A Revisão da Literatura 604
19.3 A Compilação dos Dados 605
A Decisão sobre o Conjunto de Dados Apropriado 605
A Entrada e o Armazenamento de Seus Dados 606
Inspeção, Limpeza e Sumário de Seus Dados 608
19.4 A Análise Econométrica 609
19.5 A Redação de um Ensaio Empírico 613
Introdução 613
Estrutura Conceitual (ou Teórica) 613
Modelos Econométricos e Métodos de Estimação 614
Os Dados 616
Resultados 617

Conclusões 618
Sugestões de Estilo 618
Resumo 621
Amostra de Projetos Empíricos 621
Lista de Periódicos 626
Fontes de Dados 627

Apêndice G Tabelas Estatísticas **629**

Referências Bibliográficas **637**

Glossário **645**

Índice Remissivo **667**

1 A Natureza da Econometria e dos Dados Econômicos

O Capítulo 1 examina o escopo da econometria e propõe questões gerais que resultam da aplicação dos métodos econométricos. A Seção 1.3 examina os tipos de dados usados em negócios, economia e outras ciências sociais. A Seção 1.4 faz uma discussão intuitiva das dificuldades associadas com a inferência da causalidade nas ciências sociais.

1.1 O QUE É ECONOMETRIA?

Imagine que você seja contratado pelo governo de seu Estado para avaliar a eficácia de um programa de treinamento financiado com recursos públicos. Suponha que esse programa ensine aos trabalhadores várias maneiras de como usar computadores no processo produtivo. O programa, com duração de 20 semanas, oferece cursos fora do horário do expediente. Qualquer trabalhador horista da produção pode participar, e a matrícula em todo o programa, ou em parte dele, é voluntária. Você deve determinar qual o efeito, se houver, do programa de treinamento sobre o salário-hora de cada trabalhador.

Suponha, agora, que você trabalhe para um banco de investimentos. Você deve estudar os retornos de diferentes estratégias de investimento que envolvem títulos do Tesouro dos Estados Unidos para decidir se elas estão de acordo com as teorias econômicas a elas associadas.

A tarefa de responder a tais questões pode parecer desanimadora à primeira vista. Nesse ponto, você deve ter somente uma vaga idéia de qual tipo de dados coletar. Até o fim deste curso de princípios de econometria, você provavelmente saberá como usar os métodos econométricos para avaliar, formalmente, um programa de treinamento ou testar uma simples teoria econômica.

A econometria é baseada no desenvolvimento de métodos estatísticos para estimar relações econômicas, testar teorias, avaliar e implementar políticas de governo e de negócios. A aplicação mais comum da econometria é a previsão de importantes variáveis macroeconômicas, tais como taxas de juros, taxas de inflação e produto interno bruto (PIB). Ainda que as previsões de indicadores econômicos sejam bastante visíveis e, muitas vezes, extensamente publicadas, os métodos econométricos podem ser usados em áreas econômicas que não têm nada a ver com previsões macroeconômicas. Por exemplo, estudaremos os efeitos de gastos em campanhas políticas sobre os resultados de eleições. No campo da educação, consideraremos o efeito de gastos públicos com escolas sobre o desempenho de estudantes. Além disso, aprenderemos como usar métodos econométricos para prever séries de tempo econômicas.

A econometria evoluiu como uma disciplina separada da estatística matemática, porque enfoca problemas inerentes à coleta e à análise de dados econômicos não-experimentais. **Dados não-experimentais** não são acumulados por meio de experimentos controlados de indivíduos, firmas ou seg-

mentos da economia. (Dados não-experimentais são, às vezes, chamados de **dados observacionais** para enfatizar o fato de que o pesquisador é um coletor passivo de dados.) **Dados experimentais** são freqüentemente coletados em ambientes de laboratório nas ciências naturais, mas são muito mais difíceis de serem obtidos nas ciências sociais. Embora seja possível realizar alguns experimentos sociais, geralmente é impossível conduzir os tipos de experimentos controlados necessários para avaliar questões econômicas, seja porque eles são proibitivamente dispendiosos ou moralmente repugnantes. Na Seção 1.4, apresentaremos alguns exemplos específicos das diferenças entre dados experimentais e não-experimentais.

Naturalmente, os econometristas, sempre que possível, valem-se dos estatísticos matemáticos. O método de análise da regressão múltipla é o esteio de ambos os campos, mas seu foco e sua interpretação podem diferir de forma marcante. Além disso, os economistas criaram novas técnicas para lidar com as complexidades dos dados econômicos e para testar as previsões das teorias econômicas.

1.2 PASSOS NA ANÁLISE ECONÔMICA EMPÍRICA

Os métodos econométricos são relevantes em, virtualmente, todos os ramos da economia aplicada. Eles entram em cena quando temos uma teoria econômica para testar ou quando temos em mente uma relação que apresenta alguma importância para decisões de negócios ou análises de políticas públicas. Uma análise empírica usa dados para testar uma teoria ou estimar uma relação.

Como se estrutura uma análise econômica empírica? Pode parecer óbvio, mas é importante enfatizar que o primeiro passo em qualquer análise empírica é a formulação cuidadosa da questão de interesse. Essa questão pode ser a de testar certo aspecto de uma teoria ou os efeitos de uma política governamental. Em princípio, métodos econométricos podem ser usados para responder a uma gama de questões.

Em alguns casos, especialmente aqueles que envolvem o teste de teorias econômicas, constrói-se um **modelo econômico** formal. Um modelo econômico consiste em equações matemáticas que descrevem várias relações. Os economistas são conhecidos por suas construções de modelos os quais descrevem um amplo leque de comportamentos. Por exemplo, em microeconomia intermediária, as decisões de consumo individual, sujeitas a uma restrição orçamentária, são descritas por modelos matemáticos. A premissa básica que fundamenta esses modelos é a *maximização da utilidade*. A hipótese de que os indivíduos fazem escolhas para maximizar seu bem-estar, sujeitas às restrições de recursos, oferece-nos um arcabouço muito poderoso para criar modelos econômicos tratáveis e fazer previsões bem definidas. No contexto das decisões de consumo, a maximização da utilidade leva a um conjunto de *equações de demanda*. Em uma equação de demanda, a quantidade demandada de cada produto depende do seu próprio preço, do preço dos bens substitutos e complementares, da renda do consumidor e das características individuais que influem no gosto. Essas equações podem formar a base de uma análise econométrica da demanda do consumidor.

Os economistas têm usado ferramentas econômicas básicas, tais como o arcabouço da maximização da utilidade, para explicar comportamentos que, à primeira vista, podem parecer de natureza não econômica. Um exemplo clássico é o modelo econômico de Becker (1968) sobre o comportamento criminoso.

EXEMPLO 1.1

(Modelo Econômico do Crime)

Em um artigo inspirador, o prêmio Nobel Gary Becker postulou um arcabouço da maximização da utilidade para descrever a participação de um indivíduo no crime. Certos crimes têm recompensas econômicas evidentes, mas muitos comportamentos criminosos têm custos. O custo de oportunidade do crime impede o criminoso de participar de outras atividades, como um emprego legal. Além disso, há custos associados com a possibilidade de ser capturado, e, se condenado, há os custos associados com o cumprimento de pena. Da perspectiva de Becker, a decisão de empreender a atividade ilegal é uma decisão de alocação de recursos com os benefícios e custos das atividades concorrentes sendo considerados.

Sob hipóteses gerais podemos derivar uma equação que descreve a quantidade de tempo gasto na atividade criminosa como uma função de vários fatores. Podemos representar tal função como

$$y = f(x_1, x_2, x_3, x_4, x_5, x_6, x_7), \tag{1.1}$$

em que

y = horas gastas em atividades criminosas,

x_1 = "salário" por hora ocupada em atividade criminosa,

x_2 = salário-hora em emprego legal,

x_3 = renda de outras atividades que não o crime ou um emprego legal,

x_4 = probabilidade de ser capturado,

x_5 = probabilidade de ser condenado se capturado,

x_6 = sentença esperada se condenado, e

x_7 = idade.

Outros fatores geralmente afetam a decisão de uma pessoa de participar de atividades criminosas, mas a lista acima representa o que poderia resultar de uma análise econômica formal. Como é comum na teoria econômica, não fomos específicos sobre a função $f(\cdot)$ em (1.1). Essa função depende de uma função utilidade subjacente, raramente conhecida. Entretanto, podemos usar a teoria econômica – ou a introspecção – para prever o efeito que cada variável teria sobre as atividades criminosas. Essa é a base para uma análise econométrica das atividades criminosas individuais.

A modelagem econômica formal é, às vezes, o ponto de partida da análise empírica, porém é mais comum usar a teoria econômica de modo menos formal, ou mesmo contar inteiramente com a intuição. Você pode concordar quanto aos determinantes do comportamento criminoso que aparecem na equação (1.1) serem razoavelmente baseados no senso comum; poderíamos chegar a tal equação diretamente, sem partir da maximização da utilidade. Essa visão tem algum mérito, embora haja casos em que derivações formais geram idéias que a intuição pode ignorar.

Vejamos o exemplo de uma equação que foi derivada por meio de um raciocínio um tanto informal.

EXEMPLO 1.2

(Treinamento e Produtividade do Trabalhador)

Considere o problema proposto no início da Seção 1.1. Um economista especializado em trabalho gostaria de examinar os efeitos do treinamento sobre a produtividade do trabalhador. Nesse caso, há pouca necessidade de teoria econômica formal. Um entendimento econômico básico é suficiente para perceber que fatores tais como educação, experiência e treinamento influenciam a produtividade do trabalhador. Os economistas também estão bem cientes de que os trabalhadores são pagos de acordo com sua produtividade. Esse raciocínio simples leva a um modelo tal que

$$salárioh = f(educ, exper, treina), \tag{1.2}$$

em que *salárioh* é o salário-hora, *educ* representa os anos de educação formal, *exper* refere-se aos anos de experiência no mercado de trabalho e *treina* corresponde a semanas ocupadas em treinamento. Novamente, outros fatores geralmente influenciam a taxa de salário, mas (1.2) captura a essência do problema.

Após especificarmos um modelo econômico, precisamos voltar ao que chamamos de modelo econométrico. Visto que trabalharemos com modelos econométricos ao longo deste texto, é importante saber como eles se relacionam com os modelos econômicos. Considere a equação (1.1) como exemplo. A forma da função $f(\cdot)$ deve ser especificada antes de podermos empreender uma análise econométrica. Uma segunda questão concernente a (1.1) é como lidar com variáveis que não podem ser razoavelmente observadas. Por exemplo, considere o "salário" que uma pessoa pode receber na atividade criminosa. Em princípio, tal quantidade é bem-definida, mas poderia ser difícil, se não impossível, observar o "salário" para um determinado indivíduo. Mesmo variáveis como a probabilidade de ser preso não podem ser obtidas de modo realista para um determinado indivíduo, mas pelo menos podemos observar estatísticas de detenção relevantes e derivar uma variável que se aproxime da probabilidade de prisão. Muitos outros fatores que não podemos listar, nem mesmo observar, afetam o comportamento criminoso, mas devemos de algum modo considerá-los.

As ambigüidades inerentes ao modelo econômico do crime são resolvidas ao se especificar um modelo econométrico particular, tal como:

$$crime = \beta_0 + \beta_1 salário_m + \beta_2 outrenda + \beta_3 freqpris + \beta_4 freqcond + \beta_5 sentmed + \beta_6 idade + u, \tag{1.3}$$

em que *crime* é alguma medida de freqüência da atividade criminosa, $salário_m$ é o salário que poderia ser ganho em um emprego legal, *outrenda* é a renda de outras fontes (ativos, herança etc.), *freqpris* é a freqüência de prisões por infrações anteriores (para aproximar a probabilidade de detenção), *freqcond* é a freqüência de condenações e *sentmed* é a duração média da sentença após as condenações. A escolha dessas variáveis é determinada pela teoria econômica e por considerações sobre os dados. O termo u contém fatores não observados, tais como o "salário" da atividade criminosa, o caráter moral, a formação da família e erros na mensuração de coisas como a atividade criminosa e a probabilidade de de tenção. Podemos adicionar variáveis de formação da família ao modelo, tais como o número de irmãos, a educação dos pais, e assim por diante, mas nunca poderemos eliminar u inteiramente. De fato, lidar com esse *termo de erro* ou *termo de disturbância* é, talvez, o componente mais importante de qualquer análise econométrica.

As constantes β_0, β_1, ..., β_6 são os *parâmetros* do modelo econométrico e descrevem as direções e as influências da relação entre *crime* e os fatores usados para determinar *crime* no modelo.

Um modelo econométrico completo para o Exemplo 1.2 poderia ser

$$salário = \beta_0 + \beta_1 edu + \beta_2 exper + \beta_3 treina + u, \qquad \textbf{(1.4)}$$

em que o termo u contém fatores tais como "aptidão inata", qualidade da educação, formação da família e uma miríade de outros fatores que podem influenciar o salário de uma pessoa. Se estivermos especialmente interessados nos efeitos do treinamento de trabalho, então β_3 é o parâmetro de interesse.

Na maioria dos casos, a análise econométrica começa pela especificação de um modelo econométrico, sem consideração de detalhes da criação do modelo. Geralmente seguimos essa abordagem, pois, em grande parte, a derivação cuidadosa de algo como o modelo econômico do crime toma muito tempo e pode nos levar para algumas áreas especializadas e freqüentemente difíceis da teoria econômica. O raciocínio econômico desempenhará um papel importante em nossos exemplos, e incorporaremos toda teoria econômica subjacente na especificação do modelo econométrico. No modelo econômico do exemplo do crime, começaríamos com um modelo econométrico tal como (1.3) e usaríamos o raciocínio econômico e o senso comum como guias para escolher as variáveis. Embora essa abordagem perca algumas das profusões da análise econômica, ela é comum e efetivamente aplicada por pesquisadores cautelosos.

Visto que um modelo econométrico tal como (1.3) ou (1.4) tenha sido especificado, várias *hipóteses* de interesse podem ser formuladas em termos dos parâmetros desconhecidos. Por exemplo, na equação (1.3), poderíamos levantar a hipótese de que $salário_m$, o salário que poderia ser ganho no emprego legal, não tem efeito sobre o comportamento criminoso. No contexto desse modelo econométrico específico, a hipótese é equivalente a $\beta_1 = 0$.

Uma análise empírica, por definição, requer dados. Após os dados sobre as variáveis relevantes terem sido coletados, os métodos econométricos são usados para estimar os parâmetros do modelo econométrico e para, formalmente, testar as hipóteses de interesse. Em alguns casos, o modelo econométrico é usado para fazer previsões com a finalidade de testar de uma teoria a estudo do impacto de uma política.

Como a coleta de dados é muito importante em trabalhos empíricos, a Seção 1.3 descreverá os tipos de dados com os quais, provavelmente, nos defrontaremos.

1.3 A ESTRUTURA DOS DADOS ECONÔMICOS

Os dados econômicos apresentam-se em uma variedade de tipos. Embora alguns métodos econométricos possam ser aplicados com pouca ou nenhuma modificação para muitos tipos diferentes de informações, as características especiais de alguns dados devem ser consideradas ou deveriam ser exploradas. Descreveremos a seguir as estruturas de dados mais importantes encontradas nos trabalhos aplicados.

Dados de Corte Transversal

Um **conjunto de dados de corte transversal** consiste em uma amostra de indivíduos, consumidores, empresas, cidades, estados, países ou uma variedade de outras unidades, tomada em um determinado ponto no tempo. Às vezes, os dados de todas as unidades não correspondem precisamente ao mesmo período. Por exemplo, muitas famílias podem ser pesquisadas durante diferentes semanas de um ano. Em uma aná-

lise pura de dados de corte transversal, ignoraríamos, na coleta de dados, quaisquer diferenças de tempo não importantes. Se o conjunto de famílias fosse pesquisado durante diferentes semanas do mesmo ano, ainda veríamos isso como um conjunto de dados de corte transversal.

Uma importante característica dos dados de corte transversal é que não podemos, freqüentemente, assumir que eles foram obtidos por amostragem aleatória da população subjacente. Por exemplo, se obtemos informações sobre salários, educação, experiência e outras características ao extrair aleatoriamente 500 pessoas de uma população de trabalhadores, teremos uma amostra aleatória da população de todas as pessoas que trabalham. A amostragem aleatória, matéria aprendida nos cursos introdutórios de estatística, simplifica a análise de dados de corte transversal. Uma revisão sobre amostragem aleatória aparece no Apêndice C disponível em www.thomsonlearning.com.br, na página deste livro.

Algumas vezes, a amostragem aleatória não é apropriada como uma hipótese para analisar dados de corte transversal. Por exemplo, suponha que estejamos interessados em estudar fatores que influenciam na acumulação de riqueza das famílias. Podemos estudar uma amostra aleatória de famílias, mas algumas talvez se recusem a relatar suas riquezas. Se, por exemplo, for menos provável que famílias mais ricas revelem sua riqueza, a amostra resultante sobre a riqueza não é uma amostra aleatória extraída da população de todas as famílias. Este é um exemplo de um problema de seleção amostral, um tópico avançado que discutiremos no Capítulo 17.

Outra violação da amostragem aleatória ocorre quando construímos uma amostra a partir de unidades grandes relacionadas à população, em especial a unidades geográficas. O problema provável em tais casos é que a população não é suficientemente grande para se supor, de maneira razoável, que as observações são extrações independentes. Por exemplo, se queremos explicar novas atividades de negócios entre estados, como função de taxas de salários, preços de energia, alíquotas de impostos, serviços prestados, qualidade da força de trabalho e outras características estaduais, é improvável que as atividades de negócios em um Estado próximo a outro sejam independentes. Isso revela que os métodos econométricos que discutimos funcionam, de fato, em tais situações, mas algumas vezes necessitam ser refinados. Na maioria dos casos, ignoraremos as complexidades que surgem ao analisar tais situações e trataremos esses problemas dentro do arcabouço da amostragem aleatória, mesmo quando não for tecnicamente correto fazê-lo.

Os dados de corte transversal são amplamente usados em economia e em outras ciências sociais. Em economia, a análise de dados de corte transversal está intimamente alinhada com campos da microeconomia aplicada, tais como economia do trabalho, finanças públicas estaduais e locais, organização industrial, economia urbana, demografia e economia da saúde. Dados sobre indivíduos, famílias, empresas e cidades em um determinado ponto do tempo são importantes para testar hipóteses microeconômicas e avaliar políticas governamentais.

Para a análise econométrica, os dados de corte transversal usados podem ser representados e armazenados em computadores. A Tabela 1.1 contém, de forma abreviada, um conjunto de dados de corte transversal para o ano de 1976, de 526 trabalhadores. (Esse é um subconjunto dos dados do arquivo WAGE1.RAW*.) As variáveis incluem *salárioh* (salário por hora), *educ* (anos de educação formal), *exper* (anos de experiência no mercado de trabalho), *feminino* (indicador de gênero) e *casado* (estado civil). Estas duas últimas variáveis são binárias (zero-um) por natureza, e servem para indicar características qualitativas dos indivíduos. (A pessoa é do sexo feminino ou não; a pessoa é casada ou não.) Falaremos mais sobre variáveis binárias no Capítulo 7 e seguintes.

---

* NRT: Todos os arquivos mencionados no texto têm a designação "*.RAW", mas no banco de dados os arquivos são planilhas em Excel (".XLS"), e os arquivos de trabalho de programas econométricos (como "*.WF1", do Eviews©). Portanto, a designação "*.RAW" é genérica e dá um significado de "*.matéria-prima" para aplicações e exercícios.

Tabela 1.1

Conjunto de Dados de Corte Transversal sobre Salários e outras Características Individuais

nobsi	*salárioh*	*educ*	*exper*	*feminino*	*casado*
1	3,10	11	2	1	0
2	3,24	12	22	1	1
3	3,00	11	2	0	0
4	6,00	8	44	0	1
5	5,30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11,56	16	5	0	1
526	3,50	14	5	1	0

A variável *nobsi* na Tabela 1.1 é o número da observação atribuído a cada indivíduo na amostra. Diferentemente das outras variáveis, ela não é uma característica do indivíduo. Todos os programas econométricos e estatísticos atribuem a cada unidade um número de observação. A intuição deveria dizer-lhe que, para dados como os da Tabela 1.1, não importa qual pessoa é classificada como observação um, qual pessoa é designada pela observação dois, e assim por diante. O fato de que a ordenação dos dados não importa para a análise econométrica é uma característica fundamental dos conjuntos de dados de corte transversal obtidos a partir da amostragem aleatória.

Às vezes, variáveis diferentes correspondem a diferentes períodos nos conjuntos de dados de corte transversal. Por exemplo, a fim de determinar os efeitos de políticas governamentais sobre o crescimento econômico de longo prazo, os economistas têm estudado a relação entre crescimento do PIB *per capita* real ao longo de certo período (digamos, 1960 a 1985) e variáveis determinadas, em parte, pela política governamental em 1960 (consumo do governo como percentagem do PIB e taxas de ensino médio de adultos). Tais conjuntos de dados poderiam ser representados como na Tabela 1.2, a qual constitui parte do conjunto de dados usados no estudo de De Long e Summers (1991) sobre as taxas de crescimento entre países.

A variável *cpibpcr* representa o crescimento médio do PIB *per capita* real ao longo do período 1960 a 1985. O fato de que *consgov60* (consumo do governo como percentagem do PIB) e *second60* (percentagem da população adulta com ensino médio) correspondem ao ano de 1960, enquanto *cpibpcr* é o crescimento médio ao longo do período 1960 a 1985, não leva a quaisquer problemas especiais ao tratar essas informações como um conjunto de dados de corte transversal. As observações estão ordenadas alfabeticamente por país, mas essa ordenação não afeta em nada qualquer análise subseqüente.

Tabela 1.2

Conjunto de Dados sobre Taxas de Crescimento Econômico e Características de Países

nobsp	*país*	*cpibpcr*	*consgov60*	*second60*
1	Argentina	0,89	9	32
2	Áustria	3,32	16	50
3	Bélgica	2,56	13	69
4	Bolívia	1,24	18	12
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
61	Zimbábue	2,30	17	6

Dados de Séries de Tempo

Um **conjunto de dados de séries de tempo** consiste em observações sobre uma variável ou muitas variáveis ao longo do tempo. Exemplos de dados de séries temporais incluem preços de ações, oferta de moeda, índice de preços ao consumidor, produto interno bruto, taxas anuais de homicídios e números de vendas de automóveis. Como eventos passados podem influenciar eventos futuros, e como, nas ciências sociais, as defasagens do comportamento são prevalecentes, o tempo é uma dimensão importante em um conjunto de dados de séries de tempo. Diferentemente do arranjo dos dados de corte transversal, a ordenação cronológica das observações em uma série de tempo transmite informações potencialmente importantes.

Uma característica essencial dos dados de séries de tempo que torna mais difícil a análise do que os dados de corte transversal é o fato de que raramente é possível assumir (se é que é possível) que as observações econômicas são independentes ao longo do tempo. A maioria das séries de tempo econômicas, bem como de outras séries de tempo, está relacionada – muitas vezes fortemente relacionada – com seus históricos recentes. Por exemplo, saber algo sobre o produto interno bruto do último trimestre nos diz muito sobre a provável variação do PIB durante este trimestre, visto que o PIB tende a permanecer razoavelmente estável de um trimestre para o próximo. Embora muitos procedimentos econométricos possam ser usados tanto com dados de corte transversal como com dados de séries de tempo, outros pontos podem ser considerados para especificar, apropriadamente, os modelos econométricos que usam dados de séries de tempo. Além disso, as modificações e embelezamentos das técnicas econométricas comuns foram desenvolvidas com a finalidade de considerar e explorar a natureza dependente das séries de tempo e para tratar de outras questões, tal como o fato de que algumas variáveis econômicas tendem a exibir claras tendências ao longo do tempo.

Outra característica dos dados de séries de tempo que pode requerer atenção especial é a **freqüência dos dados**, na qual eles são coletados. Em economia, as freqüências mais comuns são: diária, semanal, mensal, trimestral e anual. Os preços de ações são registrados em intervalos diários (excluindo sábados e domingos). A oferta de moeda na economia dos Estados Unidos é informada semanalmente. Muitas séries macroeconômicas são tabuladas mensalmente, incluindo as taxas de inflação e desemprego. Outras séries macroeconômicas são registradas menos freqüentemente, como a cada três meses (todo trimestre). O produto interno bruto é um exemplo importante de uma série trimestral. Outras séries de tempo, como as taxas de mortalidade infantil dos estados norte-americanos, estão disponíveis somente em bases anuais.

Muitas séries de tempo econômicas, sejam semanais, mensais ou trimestrais, exibem um forte padrão sazonal, o qual pode ser um importante fator na análise de séries de tempo. Por exemplo, dados mensais sobre o início da construção de moradias se diferenciam entre os meses simplesmente devido a mudanças das condições climáticas. Aprenderemos como trabalhar com séries de tempo no Capítulo 10.

A Tabela 1.3 contém um conjunto de dados de séries de tempo, obtido de um artigo de Castillo-Freeman e Freeman (1992), sobre os efeitos do salário mínimo em Porto Rico. O ano mais antigo no conjunto de dados é a primeira observação, e o ano mais recente disponível é a última observação. Quando os métodos econométricos são utilizados para analisar dados de séries de tempo, os dados devem ser armazenados em ordem cronológica.

Tabela 1.3

Salário Mínimo, Desemprego e Dados Relacionados para Porto Rico

nobsa	*ano*	*minmed*	*cobmed*	*desemp*	*pnb*
1	1950	0,20	20,1	15,4	878,7
2	1951	0,21	20,7	16,0	925,0
3	1952	0,23	22,6	14,8	1.015,9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
37	1986	3,35	58,1	18,9	4.281,6
38	1987	3,35	58,2	16,8	4.496,7

A variável *minmed* se refere ao salário mínimo médio no ano, *cobmed* é a taxa de cobertura média (o percentual de trabalhadores cobertos pela lei de salário mínimo), *desemp* é a taxa de desemprego e *pnb* é o produto nacional bruto. Usaremos esses dados mais adiante em uma análise de séries de tempo do efeito do salário mínimo sobre o emprego.

Cortes Transversais Agrupados

Alguns conjuntos de dados têm tanto características de corte transversal quanto de séries de tempo. Por exemplo, suponha que dois estudos sobre famílias sejam realizados nos Estados Unidos com dados de corte transversal, um em 1985 e outro em 1990. Em 1985, uma amostra aleatória de famílias é pesquisada para variáveis tais como renda, poupança, tamanho da família, e assim por diante. Em 1990, uma *nova* amostra aleatória de famílias é extraída usando as mesmas questões da pesquisa. A fim de aumentar nosso tamanho de amostra, podemos formar um **corte transversal agrupado** ao combinar os dois anos.

Agrupar cortes transversais de diferentes anos é, freqüentemente, um modo eficaz de analisar os efeitos de uma nova política de governo. A idéia é coletar dados de anos anteriores e posteriores a uma importante mudança de política governamental. Como exemplo, considere o seguinte conjunto de dados sobre os preços da moradia coletados em 1993 e 1995 nos Estados Unidos, quando houve uma redução nos impostos sobre a propriedade em 1994. Suponha que tenhamos dados sobre 250 residências para 1993 e sobre 270 para 1995. Um modo de armazenar tais dados é apresentado na Tabela 1.4.

Tabela 1.4

Cortes-Transversais Agrupados: Dois Anos de Preços de Moradias

nobsm	*ano*	*preçoc*	*imppro*	*arquad*	*qtdorm*	*banhos*
1	1993	85.500	42	1.600	3	2,0
2	1993	67.300	36	1.440	3	2,5
3	1993	134.000	38	2.000	4	2,5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
250	1993	243.600	41	2.600	4	3,0
251	1995	65.000	16	1.250	2	1,0
252	1995	182.400	20	2.200	4	2,0
253	1995	97.500	15	1.540	3	2,0
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
520	1995	57.200	16	1.100	2	1,5

As observações 1 a 250 correspondem às residências vendidas em 1993, e as observações 251 a 520 correspondem às 270 residências vendidas em 1995. Embora a ordem na qual armazenamos os dados não se revele crucial, não se esqueça de que o ano de cada observação é, geralmente, muito importante. Essa é a razão de introduzirmos *ano* como uma variável separada.

A análise de um corte transversal agrupado é muito parecida com a de um corte transversal padrão, exceto pelo fato de que precisamos, freqüentemente, considerar diferenças periódicas das variáveis ao longo do tempo. De fato, além de aumentar o tamanho da amostra, a característica de uma análise de corte transversal agrupada é, freqüentemente, ver como uma relação fundamental mudou ao longo do tempo.

Dados de Painel ou Longitudinais

Um conjunto de **dados de painel** (ou dados longitudinais) consiste em uma série de tempo para *cada* membro do corte transversal do conjunto de dados. Como exemplo, suponha que tenhamos o histórico de salário, educação e emprego para um conjunto de indivíduos ao longo de um período de dez anos, ou que possamos coletar informações, tais como dados de investimento e financeiros, sobre o mesmo conjunto de empresas ao longo de um período de cinco anos. Dados de painel também podem ser coletados para unidades geográficas. Por exemplo, podemos coletar dados para o mesmo conjunto de municípios dos Estados Unidos sobre fluxos de imigração, impostos, taxas de salários, gastos governamentais etc., para os anos de 1980, 1985 e 1990.

A característica essencial dos dados de painel que os distingue dos dados de corte transversal agrupado é o fato de que as *mesmas* unidades do corte transversal (indivíduos, empresas ou municípios nos exemplos anteriores) são acompanhadas ao longo de um determinado período. Os dados na Tabela 1.4 não são considerados um conjunto de dados de painel porque as residências vendidas são provavelmente diferentes em 1993 e 1995; se houver quaisquer repetições, o número é provavelmente bem pequeno para ser significante. Em contraste, a Tabela 1.5 contém um conjunto de dados de painel de dois anos sobre o crime e as estatísticas relacionadas para 150 cidades nos Estados Unidos.

Há várias características interessantes na Tabela 1.5. Primeiro, a cada cidade foi dado um número de 1 a 150. Qual cidade decidimos chamar de cidade 1, cidade 2, e assim por diante, é irrelevante. Assim como em um corte transversal puro, a ordenação no corte transversal de um conjunto de dados de painel não é importante. Poderíamos usar o nome da cidade em lugar de um número, mas é freqüentemente útil ter ambos.

Tabela 1.5

Conjunto de Dados de Painel sobre Estatísticas de Crime nas Cidades para Dois Anos

nobsc	*cidade*	*ano*	*homicds*	*população*	*desemp*	*polícia*
1	1	1986	5	350.000	8,7	440
2	1	1990	8	359.200	7,2	471
3	2	1986	2	64.300	5,4	75
4	2	1990	1	65.100	5,5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260.700	9,6	286
298	149	1990	6	245.000	9,8	334
299	150	1986	25	543.000	4,3	520
300	150	1990	32	546.200	5,2	493

Um segundo ponto é que os dois anos dos dados para a cidade 1 preenchem as duas primeiras linhas ou observações. As observações 3 e 4 correspondem à cidade 2, e assim por diante. Como cada uma das 150 cidades tem duas linhas de dados, qualquer pacote econométrico verá isso como 300 observações. Esse conjunto de dados pode ser tratado como um corte transversal agrupado, em que as mesmas cidades aparecem em cada ano. Porém, como veremos nos Capítulos 13 e 14, podemos também usar a estrutura de painel para responder a questões que não podem ser respondidas simplesmente vendo isso como um corte transversal agrupado.

Ao organizar as observações na Tabela 1.5, colocamos os dois anos dos dados de cada cidade um ao lado do outro, com o primeiro ano antecedendo o segundo em todos os casos. Apenas por questões práticas, esse é o modo preferido de se ordenar conjuntos de dados de painel. Essa organização contrasta com o modo pelo qual os cortes transversais agrupados são armazenados na Tabela 1.4. Em resu-

mo, a razão para ordenar os dados de painel como na Tabela 1.5 é que precisaremos fazer transformações dos dados para cada cidade nos dois anos.

Como os dados de painel requerem a repetição das mesmas unidades ao longo do tempo, os conjuntos de dados de painel, especialmente aqueles sobre indivíduos, famílias e empresas, são mais difíceis de se obter que os cortes transversais agrupados. Não surpreendentemente, observar as mesmas unidades ao longo do tempo traz várias vantagens sobre os dados de corte transversal ou mesmo sobre os de dados de cortes transversais agrupados. O benefício que salientaremos neste livro é que ter múltiplas observações sobre as mesmas unidades nos permite controlar certas características não observáveis dos indivíduos, firmas etc. Como veremos, o uso de mais de uma observação pode facilitar a inferência causal em situações em que inferir causalidade seria muito difícil se somente um único corte transversal estivesse disponível. Uma segunda vantagem dos dados de painel é que eles, freqüentemente, nos permitem estudar a importância das defasagens do comportamento ou o resultado de tomar decisões. Essa informação pode ser importante, visto que se pode esperar o impacto em muitas políticas públicas somente após algum tempo.

A maior parte dos livros para cursos de nível superior não contém uma discussão de métodos econométricos para dados de painel. Entretanto, os economistas agora reconhecem que algumas questões são difíceis, se não impossíveis, de serem respondidas satisfatoriamente sem dados de painel. Como você verá, podemos fazer consideráveis progressos com análises simples de dados de painel, um método que não é muito mais difícil do que trabalhar com um conjunto de dados de corte transversal padrão.

Um Comentário sobre Estruturas de Dados

A Parte 1 deste livro cobre a análise de dados de corte transversal, já que ela propõe menos conceitos e dificuldades técnicas. Ao mesmo tempo, ela ilustra muitos dos temas essenciais da análise econométrica. Usaremos os métodos e as idéias da análise de corte transversal no restante do texto.

Embora a análise econométrica de séries de tempo use muitas das mesmas ferramentas que a análise de corte transversal, ela é mais complicada devido à existência de tendência que traduz a natureza altamente persistente de muitas séries de tempo econômicas. Acredita-se que agora são considerados falhos muitos exemplos que têm sido tradicionalmente usados para ilustrar a maneira pela qual os métodos econométricos podem ser aplicados a dados de séries de tempo. Faz pouco sentido usar tais exemplos inicialmente, visto que esse hábito somente reforça uma prática econométrica insatisfatória. Portanto, postergaremos o tratamento da econometria de séries de tempo até a Parte 2, quando questões importantes concernentes a tendência, persistência, dinâmica e sazonalidade serão introduzidas.

Na Parte 3, trataremos explicitamente de cortes transversais agrupados e dados de painel. A análise de cortes transversais independentemente agrupados e a análise simples de dados de painel são ambas independentemente, extensões claras e diretas da análise pura de corte transversal. Entretanto, vamos esperar até o Capítulo 13 para tratar desses tópicos.

1.4 A CAUSALIDADE E A NOÇÃO DE *CETERIS PARIBUS* NA ANÁLISE ECONOMÉTRICA

Em muitos testes de teoria econômica, e certamente para avaliar políticas públicas, o objetivo do economista é inferir que uma variável (tal como a educação) tem um **efeito causal** sobre outra variável (tal como a produtividade do trabalhador). Encontrar simplesmente uma associação entre duas ou mais variáveis pode ser sugestivo, mas, a não ser que se possa estabelecer uma causalidade, raramente ela é convincente.

A noção de ***ceteris paribus*** – que significa "outros fatores (relevantes) permanecendo iguais" – desempenha um papel importante na análise causal. Essa idéia esteve implícita em algumas de nossas discussões anteriores, particularmente nos Exemplos 1.1 e 1.2, mas até agora não a mencionamos explicitamente.

Você provavelmente se lembra de que na economia introdutória muitas questões econômicas são *ceteris paribus* por natureza. Por exemplo, na análise da demanda do consumidor, estamos interessados em conhecer o efeito da variação do preço de um bem sobre sua quantidade demandada, enquanto todos os outros fatores – tais como renda, preços de outros bens e gostos individuais – permanecem fixos. Se outros fatores não forem mantidos fixos, não poderemos conhecer o efeito causal de uma variação do preço sobre a quantidade demandada.

Manter fixos os outros fatores também é crucial para a análise da política governamental. No exemplo do treinamento de trabalho (Exemplo 1.2), poderíamos nos interessar pelo efeito de outra semana de treinamento sobre os salários, com todos os outros componentes permanecendo iguais (em particular, educação e experiência). Se conseguirmos manter fixos todos os outros fatores relevantes e, em seguida, acharmos uma ligação entre treinamento e salários, poderemos concluir que o treinamento tem um efeito causal sobre a produtividade do trabalhador. Embora isso possa parecer simples, mesmo nesse estágio inicial deve ficar claro que, exceto em casos muito especiais, não será possível, literalmente, manter tudo o mais igual. A questão fundamental na maioria dos estudos empíricos é: foram mantidos fixos em número suficiente outros fatores, para que se possa inferir a causalidade? Raramente avalia-se um estudo econométrico sem levantar essa questão.

Em muitas aplicações sérias, o número de fatores que podem afetar a variável de interesse – tal como a atividade criminosa ou os salários – é imenso, e isolar qualquer variável particular pode parecer um esforço inútil. Entretanto, veremos no final que, quando cuidadosamente aplicados, os métodos econométricos podem simular um experimento *ceteris paribus*.

Neste ponto, não podemos ainda explicar como os métodos econométricos são usados para estimar efeitos *ceteris paribus*; desse modo, consideraremos alguns problemas que podem surgir ao se tentar inferir causalidade em economia. Não vamos usar nenhuma equação nessa discussão. Para cada exemplo, o problema de inferir causalidade desaparece se um experimento apropriado puder ser conduzido. Assim, é útil descrever como tal experimento poderia ser estruturado e observar que, em muitos casos, obter dados experimentais é impraticável. Também é de grande auxílio pensar por que os dados disponíveis às vezes não têm as principais características de um conjunto de dados experimentais.

Daqui em diante, contaremos com a compreensão intuitiva dos termos *aleatório*, *independência* e *correlação*, que devem ser familiares para quem estudou probabilidade e estatística. (Esses conceitos são revistos no Apêndice B, disponível na página do livro, no site www.thomsonlearning.com.br.) Vamos começar com um exemplo que ilustra algumas dessas questões importantes.

EXEMPLO 1.3

(Efeitos dos Fertilizantes sobre a Produção Agrícola)

Alguns dos primeiros estudos econométricos [por exemplo, Griliches (1957)] consideraram os efeitos de novos fertilizantes sobre a produção agrícola. Suponha a soja como o produto em consideração. Como a quantidade de fertilizantes é somente um fator que afeta a produção – outros fatores incluem chuva, qualidade da terra e presença de parasitas –, essa questão deve ser levantada como uma questão *ceteris paribus*. Uma maneira de determinar o efeito causal da quantidade de fertilizantes sobre a produção de soja é conduzir um experimento, que poderia incluir os seguintes passos. Escolha vários lotes de terra de um acre. Aplique diferentes quantidades de fertilizante em cada lote e, subseqüentemente, mensure a produção; isso nos dá

EXEMPLO 1.3 (continuação)

um conjunto de dados de corte transversal. Em seguida, use os métodos estatísticos (a serem introduzidos no Capítulo 2) para medir a associação entre produção de soja e quantidades de fertilizantes.

Como descrito anteriormente, isso pode não parecer um experimento muito bom, pois não dissemos nada sobre escolher lotes de terra que sejam idênticos em todos os aspectos, com exceção da quantidade de fertilizantes. De fato, escolher lotes de terra com essa característica não é exeqüível: alguns dos fatores, como a qualidade da terra, não podem ser, de fato, observados. Como sabemos que os resultados desse experimento podem ser usados para mensurar o efeito *ceteris paribus* dos fertilizantes? A resposta depende das especificidades de como as quantidades de fertilizantes são escolhidas. Se os níveis de fertilizantes são atribuídos aos lotes independentemente de outras características do lote que afetam a produção – isto é, outras características dos lotes são completamente ignoradas quando se decide sobre as quantidades de fertilizantes –, então podemos começar a fazer o que planejamos. Justificaremos essa afirmação no Capítulo 2.

O próximo exemplo é mais representativo das dificuldades que surgem ao se inferir causalidade em economia aplicada.

EXEMPLO 1.4

(Medindo o Retorno da Educação)

Os economistas especializados em trabalho e os *formuladores de políticas públicas* há muito se interessam pelo "retorno da educação". De modo um tanto informal, a questão é colocada da seguinte maneira: se uma pessoa é escolhida de uma população, e recebe um ano a mais de educação, em quanto aumentará seu salário? Assim como nos exemplos anteriores, essa é uma questão *ceteris paribus*, que implica que todos os outros fatores são mantidos fixos enquanto a pessoa recebe um ano a mais de educação.

Podemos imaginar um planejador social esquematizando um experimento para estudar essa questão, da mesma maneira que o pesquisador agrícola pode projetar um experimento para estimar os efeitos dos fertilizantes. Uma abordagem é seguir o exemplo dos fertilizantes no Exemplo 1.3: escolha um grupo de pessoas, dê aleatoriamente a cada pessoa uma quantidade de educação (algumas pessoas recebem alguns anos de estudo que equivalem ao ensino fundamental, a outras é dado uma educação que equivale ao ensino médio etc.), e, em seguida, mensure seus salários (assumindo que cada uma delas no momento trabalha). Aqui, as pessoas são como os lotes no exemplo dos fertilizantes, em que a educação desempenha o papel dos fertilizantes, e o salário, o da produção da soja. Como no Exemplo 1.3, se níveis de educação forem atribuídos independentemente de outras características que afetam a produtividade (tal como experiência e aptidão inata), uma análise que ignorar esses outros fatores produzirá resultados úteis. Uma vez mais, no Capítulo 2, faremos algum esforço para justificar essa afirmação; por ora, ela é formulada sem sustentação.

Diferentemente do exemplo fertilizante-produção, o experimento descrito no Exemplo 1.4 é inexeqüível. As questões morais – sem mencionar os custos econômicos – associadas à determinação aleatória dos níveis de educação para um grupo de indivíduos são óbvias. Além disso, não há lógica em simplesmente atribuir a alguém alguns anos de educação se tal pessoa já completou o curso superior.

Embora dados experimentais não possam ser obtidos para medir o retorno da educação, podemos certamente coletar dados não-experimentais sobre níveis de educação e salários para um grupo grande, fazendo amostras aleatórias da população de trabalhadores. Tais dados estão disponíveis em uma variedade de pesquisas usadas em economia do trabalho, mas esses conjuntos de dados têm uma característica que torna difícil estimar o retorno *ceteris paribus* da educação.

As pessoas *escolhem* seus próprios níveis de educação; portanto, os níveis de educação não são, provavelmente, determinados independentemente de todos os outros fatores que afetam os salários. Esse problema é uma característica compartilhada de muitos conjuntos de dados não-experimentais.

Um fator que afeta o salário é a experiência da força de trabalho. Visto que possuir mais educação requer, geralmente, um adiamento da entrada na força de trabalho, aqueles com mais educação têm, muitas vezes, menos experiência. Assim, em um conjunto de dados não-experimentais sobre salários e educação, provavelmente a educação está negativamente associada com uma variável fundamental que também afeta o salário. Acredita-se também que pessoas com mais aptidão inata escolham, freqüentemente, níveis de educação mais altos. Como aptidão maior leva a salários maiores, temos novamente uma correlação entre educação e um fator crucial que afeta o salário.

Os fatores omitidos no exemplo dos salários, experiência e aptidão, têm semelhança no exemplo dos fertilizantes. A experiência é, em geral, fácil de mensurar e, portanto, similar a uma variável como a chuva. A aptidão, no entanto, é algo vago e difícil de quantificar; ela é similar à qualidade da terra no exemplo dos fertilizantes. Como veremos ao longo deste livro, considerar outros fatores, como a experiência, ao estimar o efeito *ceteris paribus* de outra variável, como a educação, é algo relativamente direto e simples. Também descobriremos que considerar fatores inerentemente não observáveis, como a aptidão, é muito mais problemático. Pode-se dizer que muitos dos avanços nos métodos econométricos têm tentado lidar com fatores não observados nos modelos econométricos.

Podemos fazer um último paralelo entre os exemplos 1.3 e 1.4. Suponha que, no caso dos fertilizantes, as quantidades de fertilizantes não sejam completamente determinadas de modo aleatório. Em vez disso, o assistente que escolheu os níveis de fertilizante pensou que seria melhor colocar mais fertilizante nas áreas de terra de maior qualidade. (Os pesquisadores agrícolas devem ter uma idéia aproximada sobre quais áreas de terra têm melhor qualidade, ainda que eles não possam ser capazes de quantificar totalmente as diferenças.) Essa situação é completamente análoga à da relação estabelecida entre o nível de escolaridade e a aptidão não observada no Exemplo 1.4. Como terras melhores levam a safras maiores, e mais fertilizantes foram usados nas melhores áreas, qualquer relação observada entre produção de soja e quantidade de fertilizantes poderia ser espúria.

EXEMPLO 1.5

(O Efeito do Cumprimento da Lei sobre os Níveis de Criminalidade das Cidades)

A questão de como impedir a criminalidade está – e provavelmente continuará – entre nós há um bom tempo. Uma indagação especialmente importante sobre esse aspecto é: a presença de mais policiais nas ruas detém a criminalidade?

A questão *ceteris paribus* é fácil de formular. Se uma cidade fosse escolhida aleatoriamente e recebesse, por exemplo, dez policiais a mais, em quanto suas taxas de criminalidade cairiam? Outra maneira de formular a questão é: se duas cidades fossem, em todos os aspectos, iguais, exceto que a cidade A tivesse dez policiais a mais que a cidade B, em quanto difeririam as taxas de criminalidade das duas cidades?

Seria virtualmente impossível encontrar pares de comunidades idênticas em todos os aspectos, exceto no que respeita ao tamanho de suas forças policiais. Felizmente, a análise econométrica não requer isso. O que, de fato, precisamos saber é se os dados que podemos coletar sobre os níveis de criminalidade de uma comunidade e o tamanho de sua força policial podem ser vistos como experimentais. Podemos, certamente, imaginar um experimento verdadeiro, envolvendo um grande número de cidades, em que decidimos quantos policiais cada uma delas usará no ano seguinte.

EXEMPLO 1.5 (continuação)

Embora os policiais possam ser usados para produzir um efeito sobre o tamanho das forças policiais, certamente não podemos dizer a cada cidade quantos policiais ela deve empregar. Se, como é provável, a decisão de uma cidade sobre quantos policiais empregar estiver correlacionada com outros fatores relativos às cidades que afetam a criminalidade, os dados deverão ser vistos como não experimentais. De fato, um modo de ver esse problema é observar que as escolhas de uma cidade relativamente ao tamanho da força policial e a quantidade de crimes são *simultaneamente determinadas*. Vamos tratar explicitamente desse problema no Capítulo 16.

Os três primeiros exemplos que discutimos utilizaram com dados de corte transversal em vários níveis de agregação (por exemplo, do indivíduo ou da cidade). Os mesmos obstáculos surgem ao se inferir causalidade em problemas de séries de tempo.

EXEMPLO 1.6

(O Efeito do Salário Mínimo sobre o Desemprego)

Uma questão de política governamental importante, e talvez controversa, diz respeito ao efeito do salário mínimo sobre as taxas de desemprego para vários grupos de trabalhadores. Embora esse problema possa ser estudado dentro de uma variedade de estruturas de dados (dados de corte transversal, de séries de tempo ou de painel), os dados de séries de tempo são, freqüentemente, usados para observar efeitos agregados. Um exemplo de um conjunto de dados de séries de tempo relativo a taxas de desemprego e salários mínimos foi dado na Tabela 1.3.

A análise-padrão de oferta e demanda implica que, quando o salário mínimo cresce acima do salário de equilíbrio de mercado, há um movimento para cima ao longo da curva de demanda por trabalho e o emprego total diminui. (A oferta de trabalho excede a demanda por trabalho.) Para quantificar esse efeito, podemos estudar a relação entre emprego e salário mínimo ao longo do tempo. Além de algumas dificuldades especiais que podem surgir ao se lidar com dados de séries de tempo, há possíveis problemas com a inferência de causalidade. O salário mínimo nos Estados Unidos não é determinado individualmente. Várias forças econômicas e políticas exercem forte influência sobre o salário mínimo de qualquer ano. (O salário mínimo, uma vez determinado, fica geralmente congelado por muitos anos, a não ser que esteja indexado à inflação.) Assim, é provável que o salário mínimo esteja relacionado com outros fatores que têm efeito sobre os níveis de emprego.

Podemos imaginar o governo dos Estados Unidos conduzindo um experimento para determinar os efeitos do salário mínimo sobre o emprego (em vez de se preocupar com o bem-estar dos trabalhadores que ganham salários baixos). O salário mínimo poderia ser estabelecido aleatoriamente pelo governo a cada ano, enquanto os resultados do emprego poderiam ser tabulados. Os dados experimentais de séries de tempo resultantes poderiam, em seguida, ser analisados usando métodos econométricos razoavelmente simples. Mas esse cenário dificilmente descreveria como os salários mínimos são determinados.

Se pudéssemos controlar suficientemente outros fatores relacionados com o emprego, ainda poderíamos esperar estimar o efeito *ceteris paribus* do salário mínimo sobre o emprego. Nesse sentido, o problema seria muito similar aos exemplos anteriores de corte transversal.

Mesmo quando as teorias econômicas não são mais naturalmente descritas em termos de causalidade, elas geralmente têm previsões que podem ser testadas por meio de métodos econométricos. O exemplo seguinte demonstra essa abordagem.

EXEMPLO 1.7

(A Hipótese das Expectativas)

A *hipótese das expectativas* da economia financeira afirma que, dadas todas as informações disponíveis ao investidor no momento de investir, o retorno *esperado* de quaisquer dois investimentos é o mesmo. Por exemplo, considere dois possíveis investimentos, com um horizonte de investimento de três meses, adquiridos no mesmo momento. (1) Comprar um título do Tesouro norte-americano de três meses, com valor de face de $ 10.000, por um preço abaixo de $ 10.000; em três meses, você recebe $ 10.000. (2) Comprar um título de seis meses (a um preço abaixo de $ 10.000) e, em três meses, vendê-lo como um título de três meses. Cada investimento requer, mais ou menos, a mesma quantidade de capital inicial, mas há uma diferença importante. Para o primeiro investimento, você sabe exatamente qual é o retorno no momento da compra, porque você sabe o preço inicial do título de três meses e de seu valor de face. Isso não é verdade para o segundo investimento: embora saiba o preço do título de três meses quando o compra, você não conhece o preço pelo qual o venderá em três meses. Portanto, há incerteza com relação a esse investimento para aqueles que têm um horizonte de investimento de três meses.

Os retornos reais desses dois investimentos serão, em geral, diferentes. De acordo com a hipótese das expectativas, o retorno esperado do segundo investimento, dadas todas as informações no momento do investimento, deve-se igualar ao retorno de se adquirir um título de três meses. Essa teoria é razoavelmente fácil de testar, como veremos no Capítulo 11.

RESUMO

Neste capítulo introdutório, discutimos o propósito e o escopo da análise econométrica. A econometria é utilizada em todos os campos da economia aplicada para testar teorias econômicas, para informar o governo, principalmente os formuladores de políticas públicas e o setor privado, e para prever séries de tempo econômicas. Às vezes, um modelo econométrico é derivado de um modelo econômico formal, mas, em outros casos, os modelos econométricos são baseados em raciocínios econômicos informais e na intuição. O objetivo da análise econométrica é estimar os parâmetros do modelo e testar as hipóteses sobre esses parâmetros; os valores e os sinais dos parâmetros determinam a validade de uma teoria econômica e os efeitos de determinadas políticas públicas.

Dados de corte transversal, de séries de tempo, de cortes transversais agrupados e de painel são os tipos mais comuns de estruturas de dados usadas na econometria aplicada. Conjuntos de dados que envolvam uma dimensão temporal, como os dados de séries de tempo e de painel, requerem tratamento especial devido à correlação através do tempo de muitas séries econômicas. Outras questões, tais como tendências e sazonalidade, surgem na análise de séries de tempo, mas não na análise de dados de corte transversal.

Na Seção 1.4, discutimos as noções de *ceteris paribus* e de inferência causal. Em muitos casos, as hipóteses das ciências sociais são, por natureza, *ceteris paribus*: todos os outros fatores relevantes devem estar fixos ao se estudar a relação entre duas variáveis. Por causa da natureza não-experimental de muitos dados coletados nas ciências sociais, descobrir relações causais é muito desafiador.

Análise de Regressão com Dados de Corte Transversal

parte 1 do texto aborda a análise de regressão com dados de corte transversal. Ela se apóia na álgebra estudada nos cursos superiores e em conceitos básicos de probabilidade e estatística. Os Apêndices A, B e C contêm revisões completas sobre esses tópicos.

O Capítulo 2 tem início com o modelo de regressão linear simples, no qual explicamos uma variável em termos de outra. Embora a regressão simples não seja amplamente usada em econometria aplicada, ela é utilizada ocasionalmente e serve como um ponto de partida natural, pois sua álgebra e suas interpretações são relativamente simples.

Os capítulos 3 e 4 cobrem os fundamentos da análise de regressão múltipla, em que permitimos que mais variáveis afetem a variável que estamos tentando explicar. A regressão múltipla é ainda o método mais geralmente usado na pesquisa empírica, de modo que esses capítulos merecem atenção cuidadosa. O capítulo 3 enfatiza a álgebra do método de mínimos quadrados ordinários (MQO), estabelecendo ainda as condições necessárias para que os estimadores MQO sejam não-viesados e também os melhores estimadores lineares não-viesados. O Capítulo 4 trata do importante tópico da inferência estatística.

O Capítulo 5 discute as propriedades referentes às amostras grandes, ou assintóticas, dos estimadores MQO. Essa discussão oferece a justificativa para os procedimentos de inferência do Capítulo 4 quando os erros em um modelo de regressão não são normalmente distribuídos. O capítulo 6 cobre alguns tópicos adicionais da análise de regressão, incluindo questões avançadas sobre forma funcional, transformação dos dados, previsão e grau de ajuste da estimação. O Capítulo 7 explica como a informação qualitativa pode ser incorporada em modelos de regressão múltipla.

O Capítulo 8 ilustra como testar e corrigir o problema da heteroscedasticidade, ou variância não-constante, no termo erro. Mostramos como as estatísticas MQO usuais podem ser ajustadas e também apresentamos uma extensão do método MQO, conhecida como método dos *mínimos quadrados ponderados*, que explica diretamente as diferentes variâncias dos erros. O Capítulo 9 explora o importante problema da correlação entre o termo erro e uma ou mais das variáveis explicativas. Demonstramos como a utilização de uma variável *proxy* pode resolver o problema de variáveis omitidas. Adicionalmente, determinamos o viés e a inconsistência dos estimadores MQO na presença de certos tipos de erros de medida nas variáveis. Diversos problemas de tratamento dos dados são também discutidos, incluindo o problema dos *outliers*.

2 O Modelo de Regressão Simples

modelo de regressão simples pode ser usado para estudar a relação entre duas variáveis. Por razões que veremos adiante, o modelo de regressão simples tem limitações enquanto ferramenta geral para a análise empírica. No entanto, às vezes ele é apropriado como ferramenta empírica. Aprender como interpretar o modelo de regressão simples é uma boa prática para estudar a regressão múltipla, o que faremos nos capítulos subseqüentes.

2.1 DEFINIÇÃO DO MODELO DE REGRESSÃO SIMPLES

Grande parte da análise econométrica começa com a seguinte premissa: y e x são duas variáveis, representando alguma população, e estamos interessados em "explicar y em termos de x", ou em "estudar como y varia com variações em x". Discutimos alguns exemplos no Capítulo 1, incluindo: y é a produção de soja, e x, a quantidade de fertilizantes; y é o salário-hora, e x, anos de educação; e y é uma taxa de criminalidade em uma comunidade, e x, o número de policiais.

Ao escrever um modelo que "explicará y em termos de x", defrontamo-nos com três questões. Primeira, como nunca há uma relação exata entre duas variáveis, como consideramos outros fatores que afetam y? Segunda, qual é a relação funcional entre y e x? E terceira, como podemos estar certos de que estamos capturando uma relação *ceteris paribus* entre y e x (se esse for um objetivo desejado)?

Podemos resolver essas ambigüidades escrevendo uma equação que relaciona y a x. Uma equação simples é

$$y = \beta_0 + \beta_1 x + u. \qquad (2.1)$$

A equação (2.1), que supostamente é válida para a população de interesse, define o **modelo de regressão linear simples**. Ela também é chamada *modelo de regressão linear de duas variáveis* ou *modelo de regressão linear bivariada*, pois relaciona as duas variáveis x e y. Vamos discutir, agora, o significado de cada uma das quantidades em (2.1). (A propósito, o termo "regressão" tem origens que não são especialmente importantes para muitas das aplicações econométricas modernas, de modo que não o explicaremos aqui. Veja Stigler (1986) para uma história interessante da análise de regressão.)

Quando relacionadas por (2.1), as variáveis y e x têm vários nomes diferentes, os quais são intercambiáveis, como explicado em seguida. y é chamada a **variável dependente**, a **variável explicada**, a **variável de resposta**, a **variável prevista**, ou o **regressando**. x é chamada a **variável independente**, a **variável explicativa**, a **variável de controle**, a **variável previsora**, ou o **regressor**. (O termo

covariável também é usado para x.) Os termos "variável dependente" e "variável independente" são usados com freqüência em econometria. Mas esteja consciente de que o nome "independente" não se refere aqui à noção estatística de independência entre variáveis aleatórias (veja Apêndice B, disponível na página do livro, no site www.thomsonlearning.com.br).

Os termos variáveis "explicada" e "explicativa" são, provavelmente, os mais descritivos. "Resposta" e "controle" são muito usados nas ciências experimentais, em que a variável x está sob o controle do pesquisador. Não usaremos os termos "variável prevista" e "previsora", embora algumas vezes você os veja no texto. Nossa terminologia para a regressão simples está resumida na Tabela 2.1.

Tabela 2.1

Terminologia para a Regressão Simples

y	x
Variável Dependente	Variável Independente
Variável Explicada	Variável Explicativa
Variável de Resposta	Variável de Controle
Variável Prevista	Variável Previsora
Regressando	Regressor

A variável u, chamada de **termo erro** ou **perturbação** da relação, representa outros fatores, além de x, que afetam y. Uma análise de regressão simples trata, efetivamente, todos os fatores, além de x, que afetam y como não-observados. Você pode pensar em u, convenientemente, como representando o "não-observado".

A equação (2.1) também trata da questão da relação funcional entre y e x. Se os outros fatores em u são mantidos fixos, de modo que a variação em u é zero, $\Delta u = 0$, então x tem um efeito linear sobre y:

$$\Delta y = \beta_1 \Delta x \text{ se } \Delta u = 0. \qquad \textbf{(2.2)}$$

Assim, a variação em y é, simplesmente, β_1 multiplicado pela variação em x. Isso significa que β_1 é o **parâmetro de inclinação** da relação entre y e x, mantendo fixos os outros fatores em u; ele é de interesse fundamental em economia aplicada. O **parâmetro de intercepto** β_0 também tem seus usos, embora ele raramente seja central para uma análise.

EXEMPLO 2.1

(Produção de Soja e Fertilizantes)

Suponha que a produção de soja seja determinada pelo modelo

$$produção = \beta_0 + \beta_1 fertilizante + u, \qquad \textbf{(2.3)}$$

EXEMPLO 2.1 (continuação)

de modo que *y* = *produção* e *x* = *fertilizantes*. O pesquisador agrícola está interessado no efeito dos fertilizantes sobre a produção, mantendo outros fatores fixos. Esse efeito é dado por β_1. O termo erro *u* contém fatores como qualidade da terra, chuva etc. O coeficiente β_1 mede o efeito dos fertilizantes sobre a produção, mantendo outros fatores fixos: $\Delta produção = \beta_1 \Delta fertilizante$.

EXEMPLO 2.2

(Uma Equação Simples do Salário)

Um modelo que relaciona o salário de uma pessoa à educação observada e outros fatores não-observados é

$$salárioh = \beta_0 + \beta_1 educ + u. \qquad \textbf{(2.4)}$$

Se *salárioh* é medido em dólares por hora e *educ* corresponde a anos de educação formal, β_1 mede a variação no salário-hora dado um ano a mais de educação, mantendo todos os outros fatores fixos. Alguns desses fatores incluem experiência da força de trabalho, aptidão inata, permanência com o empregador atual, ética no trabalho e inumeráveis outras coisas.

A linearidade de (2.1) implica que uma variação de uma unidade em *x* tem o mesmo efeito sobre *y*, independentemente do valor inicial de *x*. Isso é irrealista para muitas aplicações econômicas. Por exemplo, no salário-educação, poderíamos querer considerar retornos *crescentes*: o próximo ano de educação teria, em relação ao anterior, um efeito *maior* sobre os salários. Veremos como considerar tais possibilidades na Seção 2.4.

A questão mais difícil é saber se o modelo (2.1) realmente nos permite tirar conclusões *ceteris paribus* sobre como *x* afeta *y*. Acabamos de ver, na equação (2.2), que β_1 mede, *de fato*, o efeito de *x* sobre *y*, mantendo todos os outros fatores (em *u*) fixos. Encerra-se com isso a questão da causalidade? Infelizmente, não. Como podemos esperar aprender algo, em geral, sobre o efeito *ceteris paribus* de *x* sobre *y*, mantendo outros fatores fixos, quando estamos ignorando todos aqueles outros fatores?

A Seção 2.5 mostrará que somos capazes de obter estimadores confiáveis de β_0 e β_1 de uma amostra aleatória de dados somente quando fazemos uma hipótese que restrinja a maneira de como o termo não-observável *u* está relacionado à variável explicativa *x*. Sem tal restrição, não seremos capazes de estimar o efeito *ceteris paribus*, β_1. Como *u* e *x* são variáveis aleatórias, precisamos de um conceito baseado em probabilidade.

Antes de expormos a hipótese crucial de como *x* e *u* são relacionados, podemos sempre fazer uma hipótese sobre *u*. Se o intercepto β_0 está incluído na equação, nada se perde ao assumir que o valor médio de *u* na população é zero.

Matematicamente,

$$E(u) = 0. \qquad \textbf{(2.5)}$$

A hipótese (2.5) não diz nada sobre a relação entre u e x; ela simplesmente faz uma afirmação sobre a distribuição dos fatores não-observáveis na população. Usando os exemplos anteriores como ilustração, podemos ver que a hipótese (2.5) não é muito restritiva. No Exemplo 2.1, não perdemos nada ao normalizar os fatores não-observáveis que afetam a produção de soja, tal como a qualidade da terra, para ter uma média zero na população de todos os lotes cultivados. O mesmo é verdadeiro para os fatores não-observáveis do Exemplo 2.2. Sem perda de generalidade, podemos assumir que coisas como a média da aptidão são zero na população de todas as pessoas que trabalham. Se você não está convencido, pode trabalhar com o Problema 2.2 para ver que podemos sempre redefinir o intercepto na equação (2.1) para tornar (2.5) verdadeiro.

Agora, vamos voltar à hipótese crucial concernente à u e x como são relacionados. Uma medida natural de associação entre duas variáveis aleatórias é o *coeficiente de correlação.* (Veja Apêndice B, disponível no site da Thomson, para definição e propriedades.) Se u e x são *não-correlacionados*, logo, enquanto variáveis aleatórias, não são *linearmente* relacionados. Assumir que u e x são não-relacionados requer um caminho longo para definir o sentido em que u e x deveriam ser não-correlacionados na equação (2.1). Mas isso não vai longe o suficiente, pois a correlação mede somente a dependência linear entre u e x. A correlação tem uma característica algo contra-intuitiva: é possível que u seja não-correlacionado com x e seja correlacionado com funções de x, tal como x^2. (Veja Seção B.4 para uma discussão adicional.) Essa possibilidade não é aceitável para muitos propósitos da regressão, visto que causa problemas para interpretar o modelo e para derivar propriedades estatísticas. Uma hipótese melhor envolve o *valor esperado de u, dado x.*

Como u e x são variáveis aleatórias, podemos definir a distribuição condicional de u, dado qualquer valor de x. Em particular, para qualquer x, podemos obter o valor esperado (ou médio) de u para aquela fatia da população descrita pelo valor de x. A hipótese crucial é que o valor médio de u *não* depende do valor de x. Podemos escrever isso como

$$E(u|x) = E(u) = 0, \tag{2.6}$$

em que a segunda igualdade resulta de (2.5). A primeira igualdade na equação (2.6) é a hipótese nova. Ela diz que, para qualquer valor de x, a média dos fatores não-observáveis é a mesma e, portanto, deve igualar-se ao valor médio de u na população. Quando combinamos a primeira igualdade da equação (2.6) com a hipótese (2.5), obtemos a **hipótese de média condicional zero.**

Vamos ver o que (2.6) acarreta ao exemplo do salário. Para simplificar a discussão, assuma que u seja o mesmo que aptidão inata. Então, (2.6) requer que o nível médio de aptidão seja o mesmo, independentemente dos anos de educação formal. Por exemplo, se $E(aptidão|8)$ representa a aptidão média para o grupo de todas as pessoas com oito anos de educação formal, e $E(aptidão|16)$ representa a aptidão média entre pessoas na população com 16 anos de educação formal, portanto (2.6) implica que essas médias devem ser as mesmas. De fato, o nível de aptidão média deve ser o mesmo para *todos* os níveis de educação. Se, por exemplo, entendemos que a aptidão média aumenta com os anos de educação formal, então (2.6) é falsa. (Isso aconteceria se, em média, pessoas com maior aptidão escolhessem tornar-se mais educadas.) Como não podemos observar aptidão inata, não temos um modo de saber se a aptidão média é ou não a mesma para todos os níveis de educação. Essa é uma questão que devemos resolver antes de aplicar a análise de regressão simples.

No exemplo dos fertilizantes, se as quantidades de fertilizantes são escolhidas independentemente de outras características dos lotes, então (2.6) se sustentará: a qualidade média da terra não dependerá da quantidade de fertilizantes. Entretanto, se mais fertilizantes forem usados em lotes de terra de melhor qualidade, então o valor esperado de u varia com o nível de fertilizantes, e (2.6) não se sustenta.

Suponha que a nota de um exame final (*nota*) dependa da freqüência às aulas (*freq*) e de fatores não-observados que afetam o desempenho dos estudantes (tal como a aptidão). Então:

$$nota = \beta_0 + \beta_1 freq + u. \qquad \textbf{(2.7)}$$

Em que situação você esperaria que esse modelo satisfaça (2.6)?

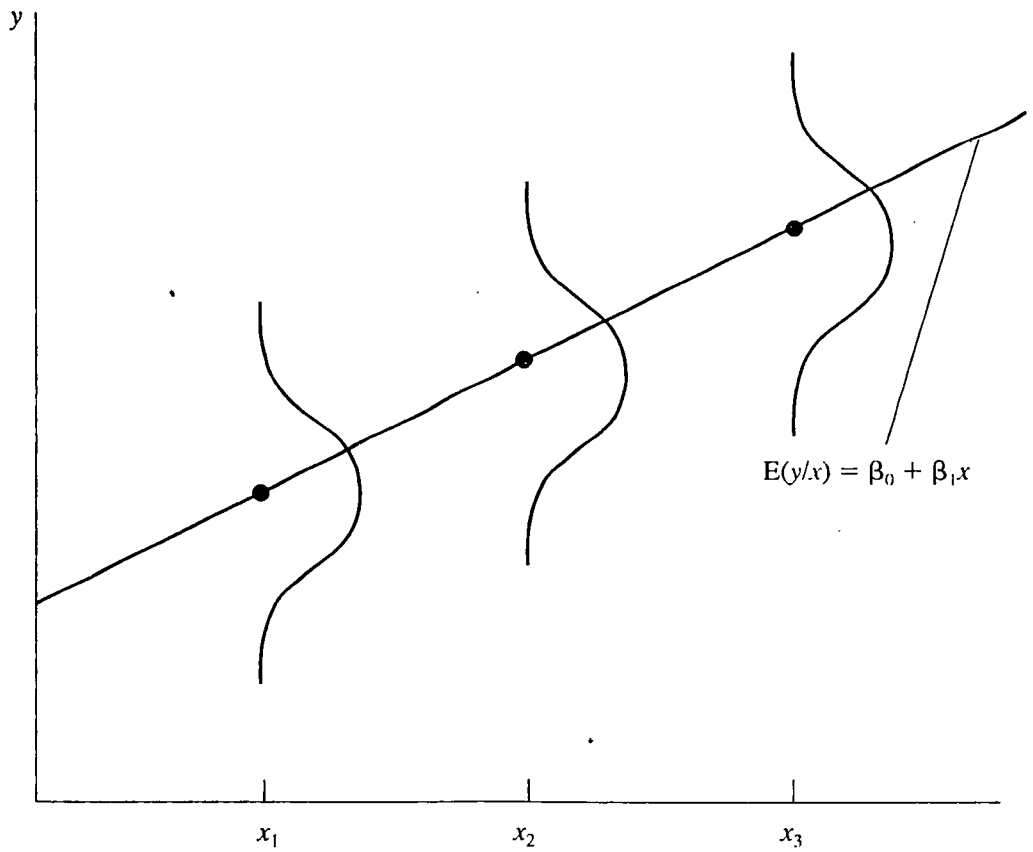
A hipótese (2.6) dá a β_1 outra interpretação que é, freqüentemente, útil. Considerando o valor esperado de (2.1) condicionado a x e usando $E(u|x) = 0$, obtém-se

$$E(y|x) = \beta_0 + \beta_1 x. \qquad \textbf{(2.8)}$$

A equação (2.8) mostra que a **função de regressão populacional** (FRP), $E(y|x)$, é uma função linear de x. A linearidade significa que o aumento de uma unidade em x faz com que o *valor esperado* de y varie segundo a magnitude de β_1. Para qualquer valor dado de x, a distribuição de y está centrada ao redor de $E(y|x)$, como ilustrado na Figura 2.1.

Figura 2.1

E(y|x) como função linear de x.



Quando (2.6) é verdadeira, é útil dividir y em dois componentes. A parte $\beta_0 + \beta_1 x$ é algumas vezes chamada a *parte sistemática* de y — isto é, a parte de y explicada por x —, e u é chamado a *parte não-sistemática*, ou a parte de y não explicada por x. Usaremos a hipótese (2.6) na próxima seção para encontrar as estimativas de β_0 e β_1. Essa hipótese também é crucial para a análise estatística na Seção 2.5.

2.2 DERIVAÇÃO DAS ESTIMATIVAS DE MÍNIMOS QUADRADOS ORDINÁRIOS

Agora que discutimos os ingredientes básicos do modelo de regressão simples, trataremos da importante questão de como estimar os parâmetros β_0 e β_1 da equação (2.1). Para tanto, necessitamos de uma amostra da população. Vamos considerar $\{(x_i, y_i): i=1, ..., n\}$ como uma amostra aleatória de tamanho n da população. Visto que esses dados vêm de (2.1), podemos escrever

$$y_i = \beta_0 + \beta_1 x_i + u_i, \tag{2.9}$$

para cada i. Aqui, u_i é o termo erro para a observação i, uma vez que ele contém todos os fatores, além de x_i, que afetam y_i.

Como um exemplo, x_i poderia ser a renda anual e y_i a poupança anual para a família i durante um determinado ano. Se coletarmos dados de 15 famílias, então $n = 15$. Um gráfico de tal conjunto de dados é dado pela Figura 2.2, juntamente com a função de regressão populacional (necessariamente fictícia).

Devemos decidir como usar esses dados, a fim de obter estimativas do intercepto e da inclinação na regressão populacional da poupança sobre a renda.

Há muitas maneiras de colocar em prática o seguinte procedimento de estimação. Usaremos (2.5) e uma importante implicação da hipótese (2.6): na população, u tem média zero e é não-correlacionado com x. Portanto, vemos que u tem valor esperado zero e que a *covariância* entre x e u é zero:

$$E(u) = 0 \tag{2.10}$$

e

$$Cov(x,u) = E(xu) = 0, \tag{2.11}$$

onde a primeira igualdade em (2.11) resulta de (2.10). (Veja Seção B.4 do Apêndice B, disponível no site da Thomson, para definição e propriedades da covariância.) Em termos das variáveis observáveis x e y e dos parâmetros desconhecidos β_0 e β_1, as equações (2.10) e (2.11) podem ser escritas como

$$E(y - \beta_0 - \beta_1 x) = 0 \tag{2.12}$$

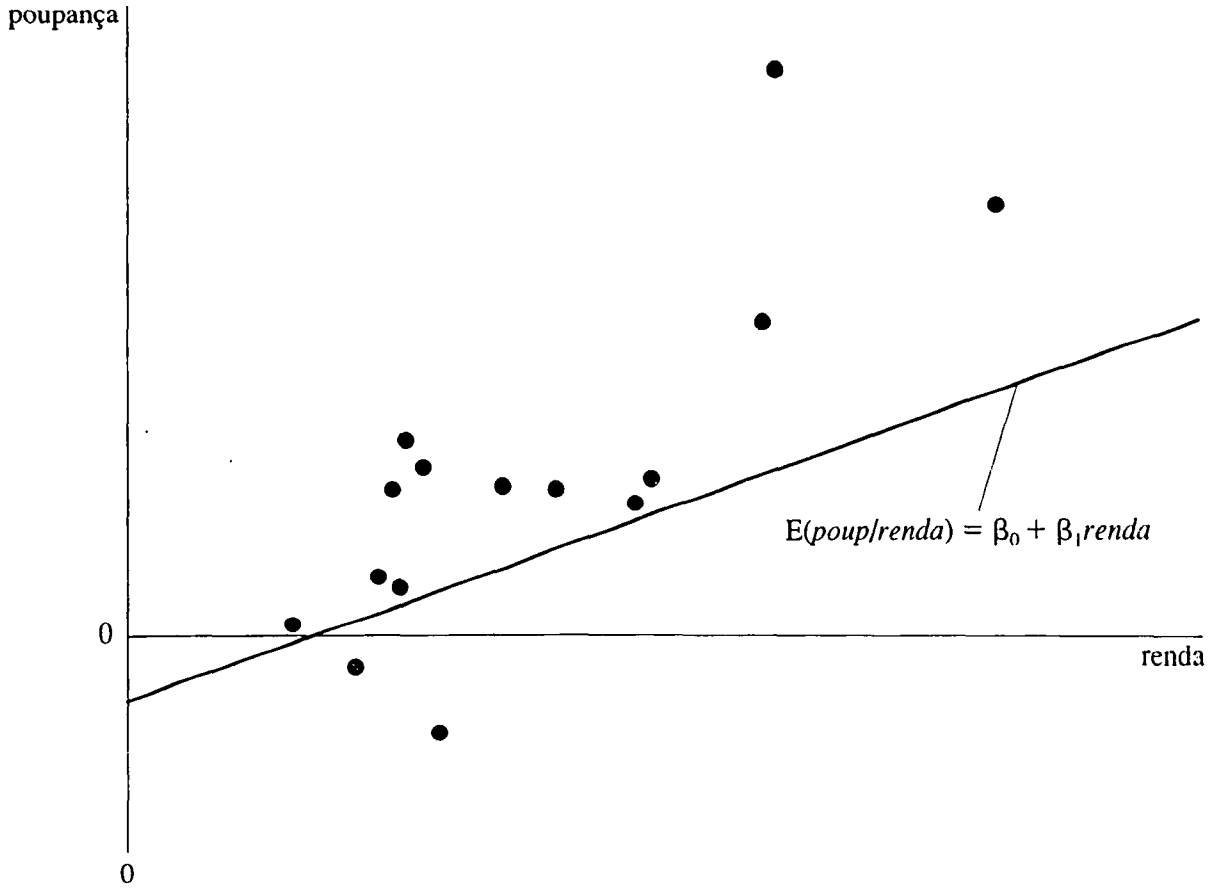
e

$$E[x(y - \beta_0 - \beta_1 x)] = 0, \tag{2.13}$$

respectivamente. As equações (2.12) e (2.13) implicam duas restrições sobre a distribuição de probabilidade conjunta de (x,y) na população. Como há dois parâmetros desconhecidos para estimar, poderíamos esperar que as equações (2.12) e (2.13) pudessem ser usadas para obter bons estimadores de β_0 e β_1. De fato, elas podem ser usadas. Dada uma amostra de dados, escolhemos as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ para resolver as equivalências *amostrais* de (2.12) e (2.13):

Figura 2.2

Gráfico da dispersão de poupança e renda de 15 famílias e a regressão populacional E(*poup*|*renda*) = $\beta_0 + \beta_1 renda$.



$$n^{-1}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \qquad \textbf{(2.14)}$$

e

$$n^{-1}\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0. \qquad \textbf{(2.15)}$$

Esse é um exemplo da abordagem do *método dos momentos* para a estimação. (Veja a Seção C.4 do Apêndice C, disponível no site da Thomson, para uma discussão das diferentes abordagens de estimação.) Essas equações podem ser resolvidas para $\hat{\beta}_0$ e $\hat{\beta}_1$.

Usando as propriedades básicas do operador somatório a partir do Apêndice A (disponível no site da Thomson), a equação (2.14) pode ser escrita como

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}, \tag{2.16}$$

em que $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ é a média amostral de y_i, e igualmente para $\bar{x}$. Essa equação nos permite escrever $\hat{\beta}_0$ em termos de $\hat{\beta}_1$, $\bar{y}$ e $\bar{x}$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{2.17}$$

Portanto, uma vez que temos a estimativa de inclinação $\hat{\beta}_1$, obtém-se diretamente a estimativa de intercepto $\hat{\beta}_0$, dados $\bar{y}$ e $\bar{x}$.

Suprimindo o n^{-1} em (2.15) (já que ele não afeta a solução) e inserindo (2.17) em (2.15), obtemos

$$\sum_{i=1}^{n} x_i[y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i] = 0$$

a qual, após rearranjo, pode ser escrita

$$\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}).$$

Das propriedades básicas do operador somatório [veja (A.7) e (A.8) disponível no site da Thomson],

$$\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \text{ e } \sum_{i=1}^{n} x_i(y_i - \bar{y}) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

Portanto, desde que

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0, \tag{2.18}$$

a inclinação estimada é

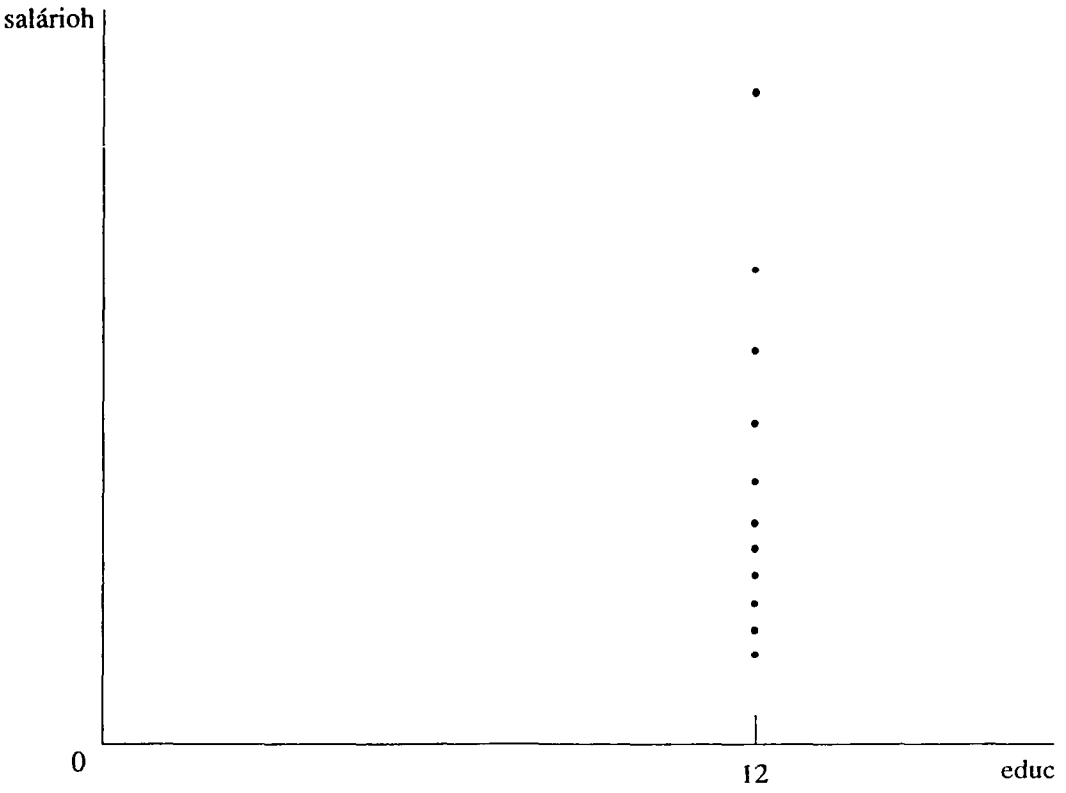
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{2.19}$$

A equação (2.19) é simplesmente a covariância amostral entre x e y, dividida pela variância amostral de x. (Veja Apêndice C, disponível no site da Thomson. Dividir tanto o numerador como o denominador por $n - 1$ não altera o resultado.) Isso faz sentido, pois β_1 é igual à covariância populacional dividida pela variância de x quando $E(u) = 0$ e $Cov(x,u) = 0$. Uma implicação imediata é que se x e y são positivamente correlacionados na amostra, então $\hat{\beta}_1$ é positivo; se x e y são negativamente correlacionados, então $\hat{\beta}_1$ é negativo.

Embora o método para obter (2.17) e (2.19) decorra de (2.6), a única hipótese necessária para se calcular as estimativas para uma amostra particular é (2.18). Mas essa raramente é uma hipótese: (2.18) é verdadeira sempre que os x_i na amostra não são todos iguais a um mesmo valor. Se (2.18) não se sustentar, então fomos infelizes em obter nossa amostra da população, ou não especificamos um problema interessante (x não varia na população). Por exemplo, se $y = salárioh$ e $x = educ$, então (2.18) não se mantém se todos na amostra têm a mesma quantidade de anos de educação formal. (Por exemplo, se todos têm o equivalente ao ensino médio concluído. Veja a Figura 2.3.) Se apenas uma pessoa tem uma quantidade diferente de anos de educação formal, então (2.18) se sustenta, e as estimativas de MQO podem ser calculadas.

Figura 2.3

Gráfico da dispersão de salários e educação, quando $educ_i = 12$ para todo i.



As estimativas dadas em (2.17) e (2.19) são chamadas de estimativas de **mínimos quadrados ordinários (MQO)** de β_0 e β_1. Para justificar esse nome, defina, para qualquer $\hat{\beta}_0$ e $\hat{\beta}_1$, um valor estimado para y quando $x = x_i$, tal como

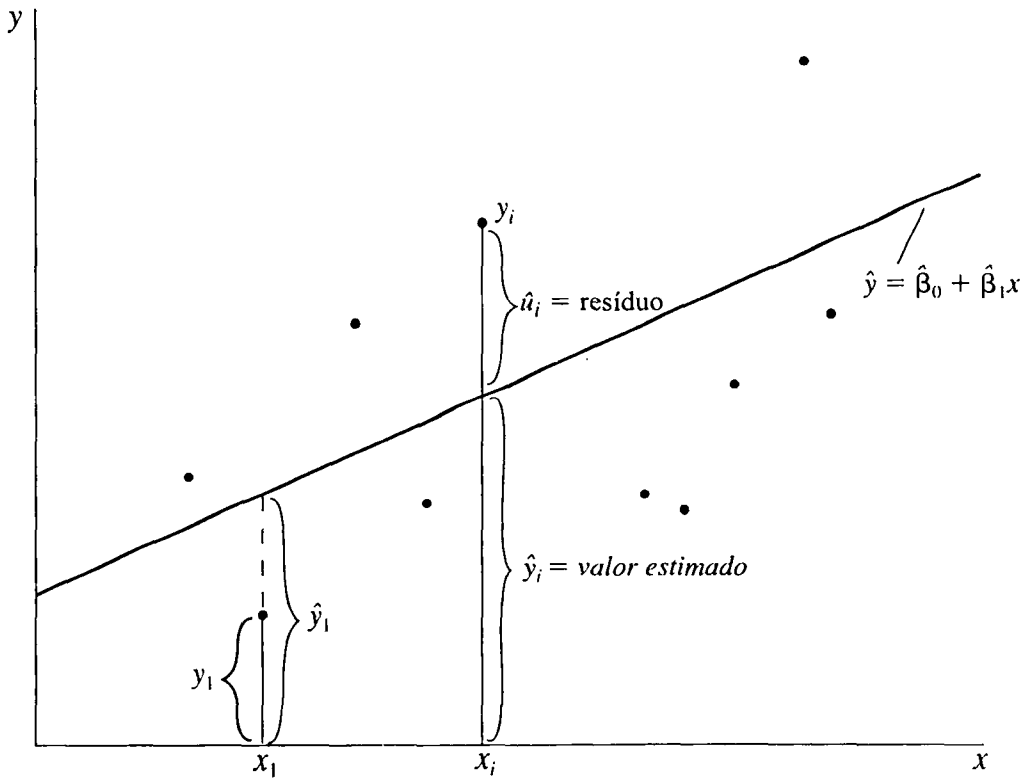
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad \textbf{(2.20)}$$

para o intercepto e a inclinação dados. Esse é o valor que prevemos para y quando $x = x_i$. Há um valor estimado para cada observação na amostra. O **resíduo** para a observação i é a diferença entre o valor verdadeiro de y_i e seu valor estimado:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i. \qquad \textbf{(2.21)}$$

Novamente, há n desses resíduos. [Eles *não* são iguais aos erros em (2.9), um ponto ao qual retornaremos na Seção 2.5.] Os valores estimados e os resíduos estão indicados na Figura 2.4.

Figura 2.4
Valores estimados e resíduos.



Agora, suponha que escolhemos $\hat{\beta}_0$ e $\hat{\beta}_1$ com a finalidade de tornar a **soma dos resíduos quadrados**,

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2, \qquad \textbf{(2.22)}$$

tão pequena quanto possível. O apêndice deste capítulo mostra que as condições necessárias para ($\hat{\beta}_0$, $\hat{\beta}_1$) minimizarem (2.22) são dadas exatamente pelas equações (2.14) e (2.15), sem n^{-1}. As equações (2.14) e (2.15) são freqüentemente chamadas de **condições de primeira ordem** para as estimativas de MQO, um termo que vem da otimização utilizada em cálculo (veja o Apêndice A, disponível no site da Thomson). De nossos cálculos anteriores, sabemos que as soluções para as condições de primeira

ordem de MQO são dadas por (2.17) e (2.19). O nome "mínimos quadrados ordinários" vem do fato de que essas estimativas minimizam a soma dos resíduos quadrados.

Quando vemos o método de mínimos quadrados ordinários como um método que minimiza a soma dos resíduos quadrados, é natural perguntar: por que não minimizar alguma outra função dos resíduos, como o valor absoluto dos resíduos? De fato, como discutiremos brevemente na Seção 9.4, minimizar a soma dos valores absolutos dos resíduos é, algumas vezes, muito útil. Mas esse procedimento também tem suas desvantagens. Primeiro, não podemos obter fórmulas para os estimadores resultantes; dado um conjunto de dados, as estimativas devem ser obtidas por rotinas de otimização numérica. Em conseqüência, a teoria estatística para estimadores que minimizam a soma dos resíduos absolutos é muito complicada. Minimizar outras funções dos resíduos, como a soma de cada resíduo elevado à quarta potência, tem desvantagens similares. (Nunca deveríamos escolher nossos estimadores para minimizar, por exemplo, a soma dos próprios resíduos, pois resíduos grandes em magnitude e com sinais opostos tendem a se cancelar.) Com o método MQO, seremos capazes de derivar, de modo relativamente fácil, inexistência de viés, consistência e outras importantes propriedades estatísticas. E mais, como as equações (2.13) e (2.14) sugerem, e como veremos na Seção 2.5, o método MQO é adequado para estimar os parâmetros que aparecem na função de média condicional (2.8).

Uma vez determinados os estimadores de intercepto e inclinação de MQO, construímos a **reta de regressão de MQO**:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{2.23}$$

em que $\hat{\beta}_0$ e $\hat{\beta}_1$ foram obtidos ao usar as equações (2.17) e (2.19). A notação $\hat{y}$ — leia-se "y chapéu" — enfatiza que os valores previstos da equação (2.23) são estimativas. O intercepto $\hat{\beta}_0$ é o valor previsto de y quando $x = 0$, embora, em alguns casos, não faça sentido considerar $x = 0$. Nestas situações, $\hat{\beta}_0$ não é, por si mesmo, muito interessante. Ao usar (2.23) para calcular os valores previstos de y para vários valores de x, devemos considerar o intercepto nos cálculos. A equação (2.23) é também chamada **função de regressão amostral (FRA)**, pois ela é a versão estimada da função de regressão populacional $E(y|x) = \beta_0 + \beta_1 x$. É importante lembrar que a FRP é algo fixo, porém desconhecido, na população. Como a FRA é obtida para uma dada amostra de dados, uma amostra nova gerará uma inclinação e um intercepto diferentes na equação (2.23).

Em muitos casos, a estimativa do coeficiente de inclinação, que podemos escrever como

$$\hat{\beta}_1 = \Delta\hat{y}/\Delta x, \tag{2.24}$$

é de interesse fundamental. Ela nos diz o quanto varia $\hat{y}$ quando x aumenta em uma unidade. Equivalentemente,

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x, \tag{2.25}$$

de modo que, dada qualquer variação em x (seja positiva ou negativa), podemos calcular a variação prevista em y.

Agora, vamos apresentar vários exemplos de regressões simples obtidas de dados reais. Em outras palavras, vamos encontrar as estimativas de intercepto e de inclinação a partir das equações (2.17) e (2.19). Como esses exemplos compreendem muitas observações, os cálculos foram feitos usando pro-

gramas econométricos. Neste ponto, não se preocupe muito em interpretar as regressões; elas não estão, necessariamente, revelando uma relação causal. Até aqui, não dissemos nada sobre as propriedades estatísticas do método MQO. Na Seção 2.5, consideraremos as propriedades depois de impormos explicitamente hipóteses sobre a equação do modelo populacional (2.1).

EXEMPLO 2.3

(Salários de Diretores Executivos e Retornos de Ações)

Para a população de diretores executivos, seja y o salário anual (*salário*) em milhares de dólares. Assim, $y = 856{,}3$ indica um salário anual de \$ 856.300, e $y = 1.452{,}6$ indica um salário de \$ 1.452.600. Seja x o retorno médio da ação sobre o patrimônio (*rma*), dos três anos anteriores, da empresa do diretor executivo. (O retorno da ação sobre o patrimônio é definido em termos de renda líquida, como uma percentagem do patrimônio comum.) Por exemplo, se $rma = 10$, então o retorno médio da ação sobre o patrimônio é de 10%.

Para estudar a relação entre essa medida do desempenho das empresas e a remuneração dos seus diretores executivos, postulamos o modelo simples

$$salário = \beta_0 + \beta_1 rma + u.$$

O parâmetro de inclinação β_1 mede a variação no salário anual, em milhares de dólares, quando o retorno da ação aumenta em um ponto percentual. Como um *rma* mais elevado é melhor para a empresa, esperamos que $\beta_1 > 0$.

O conjunto de dados do arquivo CEOSAL1.RAW contém informações sobre 209 diretores executivos para o ano de 1990; esses dados foram obtidos da revista *Business Week* (6.5.91). Na amostra, o salário médio anual é \$ 1.281.120; sendo que o menor e o maior são \$ 223.000 e \$ 14.822.000, respectivamente. O retorno médio das ações para os anos 1988, 1989 e 1990 é de 17,18%, sendo que os valores menor e maior são de 0,5 e 56,3%, respectivamente.

Usando os dados do arquivo CEOSAL1.RAW, a reta de regressão de MQO que relaciona *salário* a *rma* é

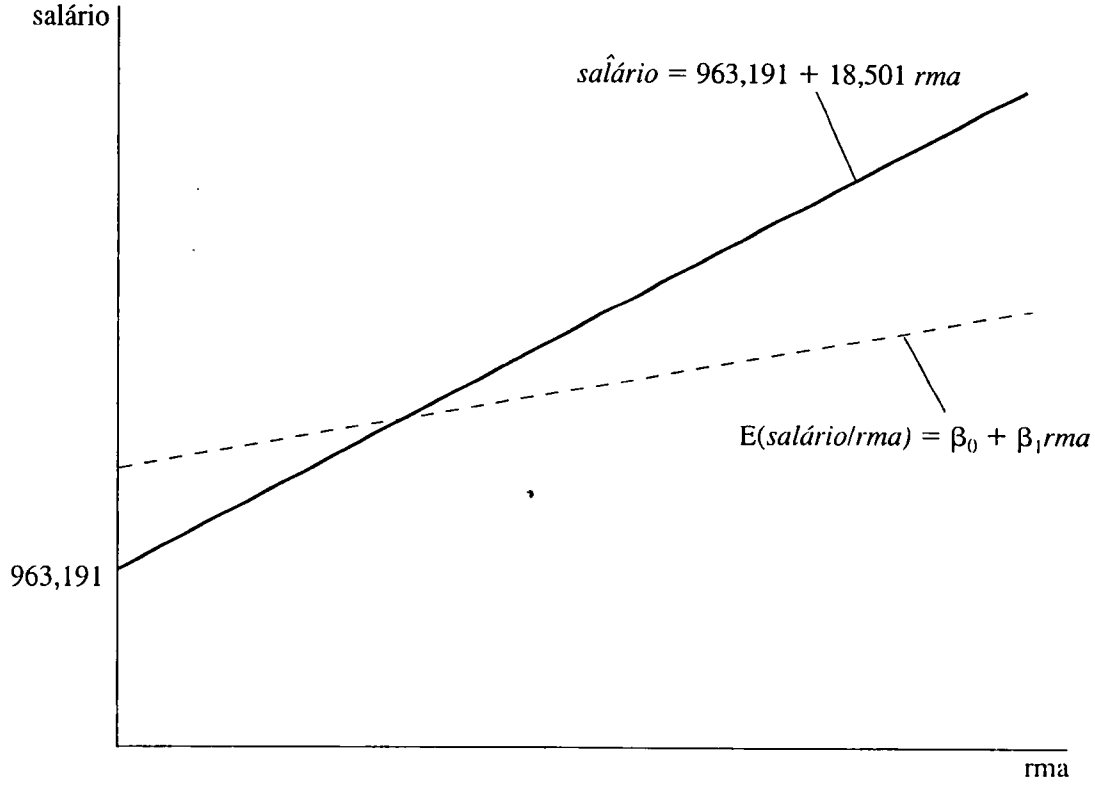
$$\widehat{salário} = 963{,}191 + 18{,}501\, rma, \qquad \textbf{(2.26)}$$

na qual as estimativas de intercepto e de inclinação foram arredondadas em três casas decimais; usamos "*salário* chapéu" para indicar que essa é uma equação estimada. Como interpretamos a equação? Primeiro, se o retorno da ação é zero, $rma = 0$, então o salário previsto é o intercepto, 963,191, que é igual a \$ 963.191, visto que *salário* é mensurado em milhares. Em seguida, podemos escrever a variação prevista no salário como uma função da variação em *rma*: $\Delta \widehat{salário} = 18{,}501(\Delta rma)$. Isso significa que se o retorno da ação aumenta um ponto percentual, $\Delta rma = 1$, então espera-se que *salário* variará cerca de 18,5, ou \$ 18.500. Como (2.26) é uma equação linear, esse valor é a variação estimada independentemente do salário inicial.

Podemos facilmente usar (2.26) para comparar salários previstos para valores diferentes de *rma*. Suponha $rma = 30$. Portanto, $\widehat{salário} = 963{,}191 + 18{,}501(30) = 1.518{,}221$, o que está pouco acima de \$ 1,5 milhão. Entretanto, isso *não* significa que um determinado diretor executivo, cuja empresa tenha um $rma = 30$, ganhe \$ 1.518.221. Muitos outros fatores afetam o salário. Essa é somente a nossa previsão a partir da reta de regressão de MQO de (2.26). A reta estimada está representada na Figura 2.5, juntamente com a função de regressão populacional $E(salário|rma)$. Nunca conheceremos a FRP; assim, não podemos dizer o quão próxima a FRA está da FRP. Outra amostra de dados levará a uma reta de regressão diferente, que pode estar, ou não, mais próxima da reta de regressão populacional.

Figura 2.5

A reta de regressão de MQO *salário* = 963,191 + 18,501 *rma* e a função de regressão populacional (desconhecida).



EXEMPLO 2.4

(Salários e Educação)

Para a população de pessoas na força de trabalho em 1976, seja $y = salárioh$, em que *salárioh* é mensurado em dólares por hora. Assim, para uma determinada pessoa, se *salárioh* = 6,75, o salário-hora é $ 6,75. Vamos chamar anos de escolaridade formal de $x = educ$; por exemplo, $educ = 12$ corresponde ao ensino médio completo (nos Estados Unidos). O salário horário médio na amostra é $ 5,90, o que equivale, de acordo com o índice de preços ao consumidor dos Estados Unidos, a $ 16,64 em dólares de 1997.

Usando os dados do arquivo WAGE1.RAW, em que $n = 526$ indivíduos, obtemos a seguinte reta de regressão de MQO (ou função de regressão amostral):

$$\widehat{salário} = -0,90 + 0,54\ educ. \qquad \textbf{(2.27)}$$

Devemos interpretar essa equação com cautela. O intercepto de −0,90 significa, literalmente, que uma pessoa sem nenhuma educação formal tem um salário-hora previsto de −90 centavos de dólar por hora. Isso, evidentemente, é tolice. Ocorre que apenas 18 pessoas na amostra de 526 têm menos que oito anos de educação formal. Conseqüentemente, não é surpreendente que a reta de regressão não faça boas previsões para

EXEMPLO 2.4 (continuação)

níveis de educação formal muito baixos. Para uma pessoa com oito anos de educação formal, o salário previsto é $\widehat{salário} = -0,90 + 0,54(8) = 3,42$, ou \$ 3,42 por hora (em dólares de 1976).

A inclinação estimada em (2.27) implica que um ano a mais de educação formal aumenta o salário horário em 54 centavos de dólar por hora. Portanto, quatro anos a mais de educação formal aumentam o salário horário previsto em 4(0,54) = 2,16, ou \$ 2,16 por hora. Esses efeitos são razoavelmente grandes. Devido à natureza linear de (2.27), outro ano de educação formal aumenta o salário na mesma quantidade, independentemente do nível inicial de educação. Na Seção 2.4, discutiremos alguns métodos que levam em consideração efeitos marginais não-constantes de nossas variáveis explicativas.

O salário horário estimado em (2.27), quando *educ* = 8, é \$ 3,42, em dólares de 1976. Qual é esse valor em dólares de 1997? (*Sugestão*: você tem informação suficiente, no Exemplo 2.4, para responder a essa questão.)

EXEMPLO 2.5

(Resultados Eleitorais e Gastos de Campanha)

O arquivo VOTE1.RAW contém dados sobre resultados eleitorais e gastos de campanha de 173 disputas entre dois partidos, para a *House of Representatives* dos Estados Unidos (equivalente a uma câmara federal). Há dois candidatos em cada disputa: A e B. Seja *votoA* a percentagem de votos recebida pelo Candidato A e *partA* a percentagem dos gastos totais de campanha que cabem ao Candidato A. Muitos outros fatores além de *partA* afetam o resultado eleitoral (incluindo a qualidade dos candidatos e os valores absolutos dos gastos de A e B). No entanto, podemos estimar um modelo de regressão simples para descobrir se gastar mais do que o concorrente implica uma percentagem maior de votos.

A equação estimada usando as 173 observações é

$$\widehat{voto}\,\mathrm{A} = 26,81 + 0,464\,partA. \qquad \textbf{(2.28)}$$

Isso significa que, se a parte dos gastos do Candidato A aumenta em um ponto percentual, o Candidato A recebe quase meio ponto percentual (0,464) a mais da votação total. Não fica claro se isso revela ou não um efeito causal, mas isso é crível. Se *partA* = 50, prevê-se que *votoA* será cerca de 50, ou metade da votação.

Em alguns casos, a análise de regressão não é usada para determinar a causalidade, mas para simplesmente observar se duas variáveis são positiva ou negativamente relacionadas, de modo muito parecido com uma análise padrão de correlação. Um exemplo disso ocorre no Problema 2.12, que pede que você use os dados de Biddle e Hamermesh (1990) referentes ao tempo que se gasta dormindo e trabalhando a fim de investigar a relação entre esses dois fatores.

No Exemplo 2.5, qual é a votação prevista para o Candidato A se *partA* = 60 (que significa 60%)? A resposta parece razoável?

Uma Nota sobre Terminologia

Em muitos casos, indicaremos a estimação de uma relação através de MQO ao escrever uma equação como (2.26), (2.27) ou (2.28). Algumas vezes, por motivo de brevidade, é útil indicar que uma regressão de MQO foi estimada sem realmente escrever a equação. Freqüentemente, indicaremos que a equação (2.23) foi obtida por MQO ao dizer que nós *rodamos* a regressão de

$$y \text{ sobre } x, \qquad \textbf{(2.29)}$$

ou simplesmente que *regredimos* y sobre x. As posições de y e x em (2.29) indicam qual é a variável dependente e qual é a variável independente: sempre regredimos a variável dependente sobre a variável independente. Para aplicações específicas, substituiremos y e x por seus nomes. Assim, para obter (2.26), regredimos *salário* sobre *rma*, ou, para obter (2.28), regredimos *votoA* sobre *partA*.

Ao usarmos essa terminologia em (2.29), sempre estaremos dizendo que planejamos estimar o intercepto, $\hat{\beta}_0$, juntamente com o coeficiente de inclinação, $\hat{\beta}_1$. Esse caso é apropriado para a maioria das aplicações. Ocasionalmente, podemos querer estimar a relação entre y e x *assumindo* que o intercepto é zero (de modo que $x = 0$ implica $\hat{y} = 0$); cobriremos esse caso, brevemente, na Seção 2.6. A não ser que seja explicitamente dito de outro modo, sempre estimaremos um intercepto juntamente com uma inclinação.

2.3 MECÂNICA DO MÉTODO MQO

Nesta seção, cobriremos algumas propriedades algébricas da reta de regressão de MQO estimada. Talvez, a melhor maneira de pensar nessas propriedades é perceber que elas são características de MQO para uma determinada amostra de dados. Elas podem ser contrastadas com as propriedades *estatísticas* de MQO, o que requer a derivação das características das distribuições amostrais dos estimadores. Discutiremos as propriedades estatísticas na Seção 2.5.

Muitas das propriedades algébricas que derivaremos parecerão triviais. No entanto, ter uma compreensão dessas propriedades ajuda-nos a entender o que acontece com as estimativas de MQO e estatísticas relacionadas quando os dados são manipulados de determinadas maneiras, como quando variam as unidades de medida das variáveis dependente e independentes.

Valores Estimados e Resíduos

Assumimos que as estimativas de intercepto e de inclinação, $\hat{\beta}_0$ e $\hat{\beta}_1$, foram obtidas de uma dada amostra de dados. Dados $\hat{\beta}_0$ e $\hat{\beta}_1$, podemos obter o valor estimado $\hat{y}_i$ para cada observação. [Isso é dado pela equação (2.20).] Por definição, cada valor estimado de $\hat{y}_i$ está sobre a reta de regressão de MQO. O resíduo de MQO associado a cada observação i, $\hat{u}_i$, é a diferença entre y_i e seu valor estimado, como dado na equação (2.21). Se $\hat{u}_i$ é positivo, a reta subestima y_i; se $\hat{u}_i$ é negativo, a reta superestima y_i. O caso ideal para a observação i é quando $\hat{u}_i = 0$, mas na maior parte dos casos *todos* os resíduos são

diferentes de zero. Em outras palavras, nenhum dos pontos dos dados deve, realmente, estar sobre a reta de MQO.

EXEMPLO 2.6

(Salário de Diretores Executivos e Retornos de Ações)

A Tabela 2.2 contém uma lista das 15 primeiras observações do conjunto de dados dos salários dos diretores executivos, juntamente com os valores estimados, chamados de *salchapéu*, e os resíduos, estimados de *uchapéu*.

Tabela 2.2

Valores Estimados e Resíduos dos 15 Primeiros Diretores Executivos

nobsd	*rma*	*salário*	*salchapéu*	*uchapéu*
1	14,1	1.095	1.224,058	−129,0581
2	10,9	1.001	1.164,854	−163,8542
3	23,5	1.122	1.397,969	−275,9692
4	5,9	578	1.072,348	−494,3484
5	13,8	1.368	1.218,508	149,4923
6	20,0	1.145	1.333,215	−188,2151
7	16,4	1.078	1.266,611	−188,6108
8	16,3	1.094	1.264,761	−170,7606
9	10,5	1.237	1.157,454	79,54626
10	26,3	833	1.449,773	−616,7726
11	25,9	567	1.442,372	−875,3721
12	26,8	933	1.459,023	−526,0231
13	14,8	1.339	1.237,009	101,9911
14	22,3	937	1.375,768	−438,7678
15	56,3	2.011	2.004,808	6,191895

Os quatro primeiros diretores executivos têm salários menores do que os previstos a partir da reta de regressão de MQO (2.26); em outras palavras, dado somente o *rma* da empresa, esses diretores executivos ganham menos do que prevemos. Como pode ser visto dos *uchapéus positivos*, o quinto diretor executivo ganha mais do que prevemos a partir da reta de regressão de MQO.

Propriedades Algébricas das Estatísticas de MQO

Há várias propriedades algébricas úteis das estimativas de MQO e das estatísticas a elas associadas. Vamos discutir as três mais importantes.

(1) A soma, e portanto a média amostral, dos resíduos de MQO, é zero. Matematicamente,

$$\sum_{i=1}^{n} \hat{u}_i = 0. \tag{2.30}$$

Essa propriedade não precisa de prova; ela resulta, imediatamente, da condição de primeira ordem de MQO (2.14), quando lembramos que os resíduos são definidos por $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. Em outras palavras, as estimativas de MQO $\hat{\beta}_0$ e $\hat{\beta}_1$ são *escolhidas* para fazer com que a soma dos resíduos seja zero (para qualquer conjunto de dados). Isso não diz nada sobre o resíduo de qualquer observação i em particular.

(2) A covariância amostral entre os regressores e os resíduos de MQO é zero. Isso resulta da condição de primeira ordem (2.15), que pode ser escrita em termos dos resíduos, como

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0. \tag{2.31}$$

A média amostral dos resíduos de MQO é zero, de modo que o lado esquerdo de (2.31) é proporcional à covariância amostral entre x_i e $\hat{u}_i$.

(3) O ponto $(\bar{x}, \bar{y})$ sempre está sobre a reta de regressão de MQO. Em outras palavras, se considerarmos a equação (2.23) e inserirmos $\bar{x}$ no lugar de x, então o valor estimado é $\bar{y}$. Isso é exatamente o que a equação (2.16) nos mostrou.

EXEMPLO 2.7

(Salários e Educação)

Para os dados do arquivo em WAGE1.RAW, o salário-hora médio da amostra é 5,90, arredondado para duas casas decimais, e a educação formal média (medida em anos) é 12,56. Se inserirmos *educ* = 12,56 na reta de regressão de MQO (2.27), obtemos $\widehat{salário}h = -0,90 + 0,54(12,56) = 5,8824$, igual a 5,9 quando arredondamos para uma casa decimal. A razão de esses números não serem exatamente os mesmos é que nós arredondamos o salário-hora e os anos de educação formal médios, assim como as estimativas de intercepto e de inclinação. Se, inicialmente, não tivéssemos arredondado nenhum dos valores, obteríamos respostas mais aproximadas, mas essa prática teria pouco efeito útil.

Ao escrever cada y_i como o seu valor estimado mais seu resíduo, temos outro modo de interpretar uma regressão de MQO. Para cada i, escreva

$$y_i = \hat{y}_i + \hat{u}_i. \tag{2.32}$$

Da propriedade (1), a média dos resíduos é zero; equivalentemente, a média amostral dos valores estimados, $\hat{y}_i$, é a mesma que a média amostral de y_i, ou $\bar{\hat{y}} = \bar{y}$. Além disso, as propriedades (1) e (2)

podem ser usadas para mostrar que a covariância amostral entre $\hat{y}_i$ e $\hat{u}_i$ é zero. Portanto, podemos ver MQO como um método que decompõe cada y_i em duas partes: um valor estimado e um resíduo. Os valores estimados e os resíduos são não-correlacionados na amostra.

Defina a **soma dos quadrados total (SQT)**, a **soma dos quadrados explicada (SQE)** e a **soma dos quadrados dos resíduos (SQR)** (também conhecida como a soma dos resíduos quadrados), como a seguir:

$$\text{SQT} \equiv \sum_{i=1}^{n}(y_i - \bar{y})^2. \tag{2.33}$$

$$\text{SQE} \equiv \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2. \tag{2.34}$$

$$\text{SQR} \equiv \sum_{i=1}^{n} \hat{u}_i^2. \tag{2.35}$$

SQT é uma medida da variação amostral total em y_i; isto é, ela mede o quão dispersos estão os y_i na amostra. Se dividirmos SQT por $n - 1$, obteremos a variância amostral de y, como discutido no Apêndice C; no site da Thomson. Semelhantemente, SQE mede a variação amostral em $\hat{y}_i$ (em que usamos o fato de que $\bar{\hat{y}} = \bar{y}$), e SQR mede a variação amostral em $\hat{u}_i$. A variação total em y pode sempre ser expressa como a soma da variação explicada e da variação não-explicada SQR. Assim,

$$\text{SQT} = \text{SQE} + \text{SQR}. \tag{2.36}$$

Provar (2.36) não é difícil; mas requer o uso de todas as propriedades do operador somatório apresentadas no Apêndice A, no site da Thomson. Escreva

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^{n}[\hat{u}_i + (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^{n}\hat{u}_i^2 + 2\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ &= \text{SQR} + 2\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y}) + \text{SQE}.\end{aligned}$$

Agora, (2.36) é válida se mostrarmos que

$$\sum_{i=1}^{n}\hat{u}_i(\hat{y}_i - \bar{y}) = 0. \tag{2.37}$$

Mas já dissemos que a covariância amostral entre os resíduos e os valores estimados é zero, e essa covariância é justamente a equação (2.37) dividida por $n - 1$. Conseqüentemente, confirmamos (2.36).

Algumas palavras de precaução sobre SQT, SQE e SQR devem ser mencionadas. Não há concordância uniforme sobre os nomes e abreviações das três quantidades definidas nas equações (2.33), (2.34) e (2.35). A soma dos quadrados total é chamada SQT ou STQ, de modo que aqui não há grandes confusões. Infelizmente, a soma dos quadrados explicada é, às vezes, chamada de "soma dos quadrados da regressão". Se a esse termo é dado sua abreviação natural, ele pode ser facilmente confundido com o termo "soma dos quadrados dos resíduos". Alguns programas econométricos referem-se à soma dos quadrados explicada como "soma dos quadrados do modelo".

Para tornar as coisas ainda piores, a soma dos quadrados dos resíduos é freqüentemente chamada de "soma dos quadrados dos erros". Esse termo é um tanto inadequado, pois, como veremos na Seção 2.5, os erros e os resíduos são quantidades diferentes. Assim, sempre chamaremos (2.35) de soma dos quadrados dos resíduos ou soma dos resíduos quadrados. Preferimos usar a abreviação SQR para representar a soma dos resíduos quadrados, pois ela é mais comum nos programas econométricos.

Grau de Ajuste

Até aqui, não apresentamos uma maneira de mensurar o quanto bem a variável explicativa ou independente, x, explica a variável dependente, y. Muitas vezes, é útil calcular um número que resume o quão bem a reta de regressão de MQO se ajusta aos dados. Na discussão seguinte, lembre-se de que assumimos estimar o intercepto com a inclinação.

Ao assumirmos que a soma dos quadrados total, SQT, não é igual a zero — o que é verdadeiro, a não ser no evento muito improvável de todos os y_i serem iguais a um mesmo valor —, podemos dividir (2.36) por SQT para obter $1 = \text{SQE/SQT} + \text{SQR/SQT}$. O ***R*-quadrado** da regressão, algumas vezes chamado coeficiente de determinação, é definido como

$$R^2 = \text{SQE/SQT} = 1 - \text{SQR/SQT}. \qquad (2.38)$$

R^2 é a razão entre a variação explicada e a variação total; assim, ele é interpretado como a *fração da variação amostral em y que é explicada por x*. A segunda equação em (2.38) fornece outra maneira de calcular R^2.

De (2.36), o valor de R^2 está sempre entre zero e um, visto que SQE não pode ser maior que SQT. Quando interpretamos R^2, usualmente o multiplicamos por 100 para transformá-lo em percentual: $100 \cdot R^2$ é a *percentagem da variação amostral em y que é explicada por x*.

Se todos os pontos dos dados estiverem sobre a mesma reta, MQO fornece um ajuste perfeito aos dados. Nesse caso, $R^2 = 1$. Um valor de R^2 quase igual a zero indica um ajuste ruim da reta de MQO: muito pouco da variação em y_i é capturado pela variação em $\hat{y}_i$ (que está sobre a reta de regressão de MQO). De fato, pode ser mostrado que R^2 é igual ao *quadrado* do coeficiente de correlação amostral entre y_i e $\hat{y}_i$. É daí que vem o termo "*R*-quadrado". (A letra R era, tradicionalmente, usada para denominar uma estimativa do coeficiente de correlação populacional, e seu uso sobreviveu na análise de regressão.)

EXEMPLO 2.8

(Salário de Diretores Executivos e Retornos de Ações)

Na regressão de salários de diretores executivos, estimamos a seguinte equação:

$$\widehat{salário} = 963{,}191 + 18{,}501\ rma \qquad \textbf{(2.39)}$$

$$n = 209, R^2 = 0{,}0132.$$

Por motivos de clareza, reproduzimos a reta de regressão de MQO e o número de observações. Usando o R-quadrado (arredondado para quatro casas decimais) apresentado para essa equação, podemos ver quanto da variação no salário é, realmente, explicada pelo retorno da ação. A resposta é: não muito. O retorno da ação da empresa explica somente 1,3% da variação nos salários dessa amostra de 209 diretores executivos. Isso significa que 98,7% das variações salariais desses diretores executivos são deixadas sem explicação. Essa falta de poder explicativo não deve ser surpreendente demais, já que muitas outras características, tanto da empresa como do diretor executivo individual, devem influenciar o salário; esses fatores estão, necessariamente, incluídos nos erros de uma análise de regressão simples.

Nas ciências sociais não são incomuns R-quadrados baixos nas equações de regressão, especialmente na análise de corte transversal. Discutiremos essa questão, de modo mais geral, sob a análise de regressão múltipla, mas vale a pena enfatizar agora que um R-quadrado aparentemente baixo não significa, necessariamente, que uma equação de regressão de MQO é inútil. Ainda, é possível que (2.39) seja uma boa estimativa da relação *ceteris paribus* entre *salário* e *rma*; se isso for verdade ou não, *não* depende diretamente da magnitude do R-quadrado. Os estudantes que estão se defrontando com econometria pela primeira vez tendem, ao avaliar equações de regressão, a pôr muito peso na magnitude do R-quadrado. Por enquanto, esteja ciente de que usar o R-quadrado como o principal padrão de medida de sucesso de uma análise econométrica pode levar a confusões.

Algumas vezes, a variável explicativa elucida uma parte substancial da variação amostral na variável dependente.

EXEMPLO 2.9

(Resultados Eleitorais e Gastos de Campanha)

Na equação de resultados eleitorais (2.28), $R^2 = 0{,}856$. Assim, a participação dos candidatos nos gastos de campanha explica mais de 85% da variação nos resultados eleitorais nessa amostra. Essa é uma explicação considerável.

2.4 UNIDADES DE MEDIDA E FORMA FUNCIONAL

Duas questões importantes em economia aplicada são: (1) entender como, ao mudar as unidades de medida das variáveis dependente e/ou independente, são afetadas as estimativas de MQO e (2) saber como incorporar, à análise de regressão, formas funcionais populares usadas em economia.

A matemática necessária para uma compreensão completa das questões sobre a forma funcional está revista no Apêndice A, disponível no site da Thomson.

Os Efeitos de Mudanças das Unidades de Medida sobre as Estatísticas de MQO

No Exemplo 2.3, escolhemos mensurar o salário anual em milhares de dólares, e o retorno das ações foi medido como uma percentagem (em vez de um decimal). É crucial saber como *salário* e *rma* são medidos nesse exemplo, a fim de dar sentido às estimativas da equação (2.39).

Devemos também saber que as estimativas de MQO mudam de maneira completamente esperada, quando as unidades de medida das variáveis dependente e independente mudam. No Exemplo 2.3, suponhamos que, em vez de medir o salário em milhares de dólares, nós o medimos em dólares. Seja *salardol* o salário em dólares (*salardol* = 845.761 seria interpretado como $ 845.761). Evidentemente, *salardol* tem uma relação simples com o salário medido em milhares de dólares: *salardol* = 1.000. *salário*. Não precisamos, realmente, computar a regressão *salardol* sobre *rma* para saber que a equação estimada é:

$$\widehat{salardol} = 963.191 + 18.501\ rma. \qquad \textbf{(2.40)}$$

Obtemos o intercepto e a inclinação em (2.40) ao, simplesmente, multiplicarmos o intercepto e o coeficiente de inclinação em (2.39) por 1.000. Isso dá às equações (2.39) e (2.40) a mesma interpretação. Olhando para (2.40), se *rma* = 0, então $\widehat{salardol}$ = 963.191, de modo que o salário previsto é $ 963.191 [o mesmo valor que obtivemos da equação (2.39)]. Além disso, se *rma* aumenta em um, então o salário previsto aumenta em $ 18.501; novamente, isso é o que concluímos de nossa análise anterior da equação (2.39).

Em geral, é fácil fazer uma idéia do que acontece às estimativas de intercepto e de inclinação quando se altera a unidade de medida da variável dependente. Se a variável dependente é multiplicada pela constante c — o que significa dizer que cada valor na amostra é multiplicado por c —, então as estimativas de MQO de intercepto e de inclinação também são multiplicadas por c. (Isso assume que nada foi alterado com respeito à variável independente.) No exemplo do salário dos diretores executivos, c = 1.000 ao passarmos de *salário* para *salardol*.

Suponha que o salário seja mensurado em centenas de dólares, em vez de milhares de dólares, e o chamemos *salarcent*. Quais serão as estimativas de intercepto e de inclinação na regressão de *salarcent* sobre *rma*?

Também podemos usar o exemplo do salário dos diretores executivos para ver o que acontece quando as unidades de medida da variável independente são mudadas. Defina *rmadec* = *rma*/100 como sendo o equivalente decimal de *rma*; assim, *rmadec* = 0,23 significa um retorno da ação de 23%. A fim de centrarmos o foco na mudança das unidades de medida da variável independente, retornaremos à nossa variável dependente original, *salário*, mensurada em milhares de dólares. Quando regredimos *salário* sobre *rmadec*, obtemos

$$\widehat{salário} = 963{,}191 + 1.850{,}1\ rmadec. \qquad \textbf{(2.41)}$$

O coeficiente de *rmadec* é 100 vezes o coeficiente de *rma* em (2.39). Isso é o que deveria ser. Variar *rma* em um ponto percentual é equivalente a $\Delta rmadec$ = 0,01. De (2.41), se $\Delta rmadec$ = 0,01, então

$\Delta\hat{salário} = 1.850{,}1(0{,}01) = 18{,}501$, que é igual ao obtido ao se usar (2.39). Observe que, ao passarmos de (2.39) para (2.41), a variável independente foi dividida por 100, e assim a estimativa de inclinação de MQO foi multiplicada por 100, preservando a interpretação da equação. Em geral, se a variável independente é dividida ou multiplicada por alguma constante diferente de zero, c, então o coeficiente de inclinação de MQO é multiplicado ou dividido por c, respectivamente.

O intercepto não mudou em (2.41), pois $rmadec = 0$ ainda corresponde a um retorno zero da ação. Em geral, mudar somente as unidades de medida da variável independente não afeta o intercepto.

Na seção anterior, definimos R-quadrado como uma medida de grau de ajuste para a regressão de MQO. Podemos também questionar o que acontece ao R^2 quando é mudada a unidade de medida da variável independente ou da variável dependente. Sem fazer nenhuma álgebra, deveríamos saber o resultado: o grau de ajuste do modelo não depende das unidades de medida de nossas variáveis. Por exemplo, a quantidade de variação no salário, explicada pelo retorno da ação, não deve depender de o salário ser medido em dólares ou em milhares de dólares, ou de o retorno da ação ser uma percentagem ou um decimal. Essa intuição pode ser verificada matematicamente: usando a definição de R^2, pode ser mostrado que R^2 é, de fato, invariante a mudanças nas unidades de y ou x.

Incorporação de Não-Linearidades na Regressão Simples

Até aqui, enfatizamos as relações *lineares* entre as variáveis dependente e independente. Como mencionamos no Capítulo 1, relações lineares não são, em geral, suficientes para todas as aplicações econômicas. Felizmente, é bastante fácil incorporar muitas não-linearidades na análise de regressão simples ao definir apropriadamente as variáveis dependente e independente. Vamos tratar aqui de duas possibilidades que freqüentemente aparecem em trabalhos aplicados.

Ao ler trabalhos aplicados nas ciências sociais, com freqüência você encontrará equações de regressão em que a variável dependente aparece na forma logarítmica. Por que isso é feito? Lembre-se do exemplo salários-educação, em que regredimos o salário-hora sobre os anos de educação formal. Obtivemos uma estimativa da inclinação de 0,54 [veja a equação (2.27)], o que significa dizer que, para cada ano adicional de educação, é previsto um aumento de 54 centavos de dólar no salário-hora. Devido à natureza linear de (2.27), 54 centavos de dólar é o aumento tanto para o primeiro ano de educação quanto para o vigésimo ano; isso pode não ser razoável.

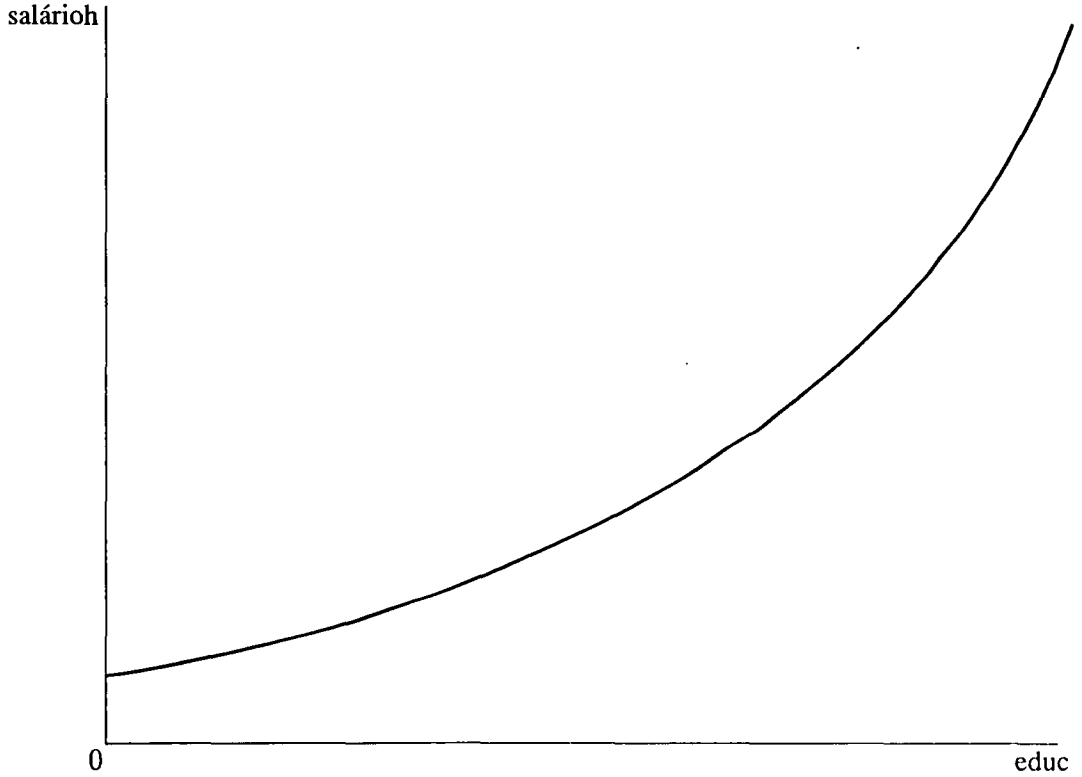
Suponha, em vez disso, que o aumento percentual no salário é o mesmo, dado um ano a mais de educação formal. O modelo (2.27) não implica um aumento percentual constante: o aumento depende do salário inicial. Um modelo que gera (aproximadamente) um efeito percentual constante é

$$\log(salárioh) = \beta_0 + \beta_1 educ + u, \qquad \textbf{(2.42)}$$

em que $\log(\cdot)$ é o logaritmo *natural*. (Veja Apêndice A, disponível no site da Thomson, para uma revisão sobre logaritmos.) Em particular, se $\Delta u = 0$, então

$$\%\Delta salárioh \approx (100 \cdot \beta_1)\Delta educ. \qquad \textbf{(2.43)}$$

Figura 2.6
salárioh = $\exp(\beta_0 + \beta_1 educ)$, *com* $\beta_1 > 0$.



Observe como multiplicamos β_1 por 100 para obter a variação percentual em *salárioh* dado um ano adicional de educação formal. Como a variação percentual em *salárioh* é a mesma para cada ano adicional de educação, a variação em *salárioh*, para um ano extra de educação formal, *aumenta* quando a educação formal aumenta; em outras palavras, (2.42) implica um retorno *crescente* da educação formal. Com a exponenciação de (2.42), podemos escrever *salárioh* = $\exp(\beta_0 + \beta_1 educ + u)$. O gráfico dessa equação aparece na Figura 2.6, com $u = 0$.

Quando se usa a regressão simples, a estimação de um modelo como (2.42) é imediata. Apenas defina a variável dependente, y, como $y = \log(salárioh)$. A variável independente é representada por $x = educ$. A mecânica de MQO é a mesma de antes: as estimativas de intercepto e de inclinação são dadas pelas fórmulas (2.17) e (2.19). Em outras palavras, obtemos $\hat{\beta}_0$ e $\hat{\beta}_1$ da regressão de MQO de log(*salárioh*) sobre *educ*.

EXEMPLO 2.10

(Uma Equação do Logaritmo dos Salários-Hora)

Utilizando os mesmos dados do Exemplo 2.4, mas usando log(*salárioh*) como a variável dependente, obtemos a seguinte relação:

$$\widehat{\log(salário)} = 0{,}584 + 0{,}083\ educ \qquad \textbf{(2.44)}$$

EXEMPLO 2.10 (continuação)

$$n = 526, R^2 = 0{,}186.$$

O coeficiente de *educ* tem uma interpretação percentual quando ele é multiplicado por 100: para cada ano adicional de educação formal, *salárioh* aumenta 8,3%. Isso é o que os economistas querem dizer quando se referem ao "retorno de um ano adicional de educação formal".

É importante relembrar que a principal razão para usar o log de *salárioh* em (2.42) é impor um efeito percentual constante da educação formal sobre *salárioh*. Uma vez obtida a equação (2.42), o log natural de *salárioh* é raramente mencionado. Em particular, *não* é correto dizer que um ano adicional de educação formal aumenta log(*salárioh*) em 8,3%.

O intercepto em (2.42) não tem muito significado, visto que ele é o log(*salárioh*) previsto quando *educ* = 0. O *R*-quadrado mostra que *educ* explica cerca de 18,6% da variação em log(*salárioh*) (não em *salário*). Finalmente, a equação (2.44) pode não capturar toda a não-linearidade da relação entre salário-hora e escolaridade formal. Se houver "efeitos-diploma", o décimo segundo ano de educação — formatura do ensino médio nos Estados Unidos — deve ser muito mais valioso que o décimo primeiro ano. No Capítulo 7 aprenderemos como lidar com esse tipo de não-linearidade.

Outro uso importante do log natural está em obter um **modelo de elasticidade constante**.

EXEMPLO 2.11

(Salários de Diretores Executivos e Vendas das Empresas)

Podemos estimar um modelo de elasticidade constante que relaciona o salário dos diretores executivos às vendas das empresas. O conjunto de dados é o mesmo utilizado no Exemplo 2.3, exceto que agora relacionamos *salário* a *vendas*. Seja *vendas* as vendas anuais das empresas, mensurada em milhões de dólares. Um modelo de elasticidade constante é

$$\widehat{\log(salário)} = \beta_0 + \beta_1 \log(vendas) + u, \qquad \textbf{(2.45)}$$

em que β_1 é a elasticidade de *salário* com respeito a *vendas*. Esse modelo está compreendido no modelo de regressão simples ao se definir a variável dependente como $y = \log(salário)$, e a variável independente como $x = \log(vendas)$. Ao estimar essa equação por MQO, temos

$$\widehat{\log(salário)} = 4{,}822 + 0{,}257\log(vendas) \qquad \textbf{(2.46)}$$

$$n = 209, R^2 = 0{,}211.$$

O coeficiente de log(*vendas*) é a elasticidade estimada de *salário* em relação a *vendas*. Ela implica que um aumento de 1% nas vendas das empresas aumenta o salário dos diretores executivos em cerca de 0,257% — a interpretação usual de uma elasticidade.

As duas formas consideradas nesta seção surgirão no restante deste texto. Tratamos aqui de modelos que contêm logaritmos naturais porque eles aparecem muito freqüentemente em trabalhos aplicados. A interpretação desses modelos não será muito diferente no caso da regressão múltipla.

É também útil observar o que acontece às estimativas de intercepto e de inclinação se mudarmos as unidades de medida da variável dependente quando ela aparece na forma logarítmica. Pelo fato de a variação da forma logarítmica aproximar-se de uma variação proporcional, faz sentido que *nada* aconteça com a inclinação. Podemos ver isso ao escrever a variável em uma nova escala como $c_1 y_i$ para cada observação i. A equação original é $\log(y_i) = \beta_0 + \beta_1 x_i + u_i$. Se adicionamos $\log(c_1)$ a ambos os lados da equação, obtemos $\log(c_1) + \log(y_i) = [\log(c_1) + \beta_0] + \beta_1 x_i + u_i$ ou $\log(c_1 y_i) = [\log(c_1) + \beta_0] + \beta_1 x_i + u_i$. (Lembre-se de que a soma dos logs é igual ao log de seus produtos, como mostrado no Apêndice A, disponível no site da Thomson.) Portanto, a inclinação ainda é β_1, mas o intercepto agora é $\log(c_1) + \beta_0$. Semelhantemente, se a variável independente for $\log(x)$, e mudarmos as unidades de medida de x antes de considerarmos o log, a inclinação permanece a mesma, mas o intercepto muda. Pediremos que você verifique essas asserções no Problema 2.9.

Finalizamos esta subseção resumindo quatro combinações de formas funcionais construídas a partir da variável original ou de seu logaritmo natural. Na Tabela 2.3, x e y representam as variáveis em suas formas originais. O modelo com y como a variável dependente e x como a variável independente é chamado modelo *nível-nível*, pois cada variável aparece em sua forma de nível. O modelo com $\log(y)$ como a variável dependente e x como a variável independente é chamado modelo *log-nível*. Não discutiremos aqui, explicitamente, o modelo *nível-log*, pois ele aparece menos freqüentemente na prática. De qualquer forma, veremos exemplos desse modelo em outros capítulos.

A última coluna na Tabela 2.3 mostra a interpretação de β_1. No modelo log-nível, $100 \cdot \beta_1$ é algumas vezes chamado **semi-elasticidade** de y em relação a x. Como mencionamos no Exemplo 2.11, no modelo log-log β_1 é a **elasticidade** de y em relação a x. A Tabela 2.3 requer um estudo cuidadoso, já que vamos, com freqüência, nos referir a ela no restante do texto.

O Significado da Regressão "Linear"

O modelo de regressão simples que estudamos neste capítulo também é chamado modelo de regressão *linear simples*. No entanto, como acabamos de ver, o modelo geral também permite certas relações *não-lineares*. Portanto, o que significa "linear" aqui? Você pode observar, ao olhar a equação (2.1), que $y = \beta_0 + \beta_1 x + u$. O importante é que essa equação é linear nos *parâmetros*, β_0 e β_1. Não há restrições de como y e x se relacionam com as variáveis explicada e explicativa originais de interesse. Como vimos nos Exemplos 2.7 e 2.8, y e x podem ser os logaritmos naturais de variáveis, e isso é muito comum em aplicações. Mas não precisamos parar aqui. Por exemplo, nada nos impede de usar a regressão simples para estimar um modelo tal como $cons = \beta_0 + \beta_1 \sqrt{rend} + u$, em que *cons* é o consumo anual e *rend* é a renda anual.

Tabela 2.3

Resumo das Formas Funcionais Envolvendo Logaritmos

Modelo	*Variável Dependente*	*Variável Independente*	*Interpretação de* β_1
nível-nível	y	x	$\Delta y = \beta_1 \Delta x$
nível-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
log-nível	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Embora a mecânica da regressão simples não dependa de como y e x são definidos, a interpretação dos coeficientes depende, realmente, de suas definições. Para que os trabalhos empíricos sejam bem-sucedidos, é muito mais importante tornar-se proficiente em interpretar coeficientes do que eficiente no cálculo de fórmulas como (2.19). Obteremos muito mais prática em como interpretar as estimativas da reta de regressão de MQO quando estudarmos a regressão múltipla.

Muitos modelos *não podem* ser considerados modelos de regressão linear, porque eles não são lineares em seus parâmetros; um exemplo é $cons = 1/(\beta_0 + \beta_1 rend) + u$. A estimação desses modelos leva-nos ao campo âmbito do *modelo de regressão não-linear*, o qual está além do escopo deste texto. Para muitas aplicações, é suficiente escolher um modelo que possa ser expresso dentro do arcabouço da regressão linear.

2.5 VALORES ESPERADOS E VARIÂNCIAS DOS ESTIMADORES DE MQO

Na Seção 2.1 definimos o modelo populacional $y = \beta_0 + \beta_1 x + u$ e afirmamos que a hipótese fundamental para que a análise de regressão simples seja útil é que o valor esperado de u, dado qualquer valor de x, seja zero. Nas seções 2.2, 2.3 e 2.4 discutimos as propriedades algébricas da estimação de MQO. Retornamos agora ao modelo populacional e estudaremos as propriedades *estatísticas* da estimação de MQO. Em outras palavras, veremos agora $\hat{\beta}_0$ e $\hat{\beta}_1$ como *estimadores* dos parâmetros β_0 e β_1 que aparecem no modelo populacional. Isso significa que estudaremos as propriedades das distribuições de $\hat{\beta}_0$ e $\hat{\beta}_1$ de diferentes amostras aleatórias da população. (O Apêndice C, disponível no site da Thomson, contém as definições de estimadores e revisões de algumas de suas principais propriedades.)

Inexistência de Viés em MQO

Vamos iniciar estabelecendo a inexistência de viés do método MQO sob um conjunto simples de hipóteses. Para referências futuras, é útil numerar essas hipóteses usando o prefixo "RLS" para regressão linear simples. A primeira hipótese define o modelo populacional.

HIPÓTESE RLS.1 (LINEAR NOS PARÂMETROS)
No modelo populacional, a variável dependente y está relacionada à variável independente x e ao erro (ou perturbação) u como

$$y = \beta_0 + \beta_1 x + u, \qquad \textbf{(2.47)}$$

em que β_0 e β_1 são os parâmetros de intercepto e de inclinação populacionais, respectivamente.

Ao especificar o modelo populacional — e para ser realista —, y, x e u são todos vistos como variáveis aleatórias. Discutimos, em alguma extensão, a interpretação desse modelo na Seção 2.1 e demos vários exemplos. Na seção anterior aprendemos que a equação (2.47) não é tão restritiva quanto inicialmente parecia; escolhendo y e x apropriadamente, podemos obter relações não-lineares interessantes (como os modelos de elasticidade constante).

Estamos interessados em usar os dados de y e x para estimar os parâmetros β_0 e, especialmente, β_1. Assumimos que nossos dados foram obtidos de uma amostra aleatória. (Veja o Apêndice C, disponível no site da Thomson, para uma revisão sobre amostragem aleatória.)

HIPÓTESE RLS.2 (AMOSTRAGEM ALEATÓRIA)
Podemos usar uma amostra aleatória de tamanho n, $\{(x_i,y_i): i = 1,2,..., n\}$, proveniente de um modelo populacional.

Em capítulos posteriores que abordam a análise de séries de tempo e problemas de seleção de amostra, teremos de dar um tratamento ao fato de a hipótese de amostragem aleatória não ser mais válida. Nem todas as amostras de corte transversal podem ser vistas como resultados de amostras aleatórias, mas muitas podem ser assim entendidas.

Podemos escrever (2.47), em termos da amostra aleatória como

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1,2,...,n, \tag{2.48}$$

em que u_i é o erro ou perturbação da observação i (por exemplo, pessoa i, empresa i, cidade i etc.). Assim, u_i contém os fatores não-observáveis da observação i que afetam y_i. Os u_i não devem ser confundidos com os resíduos, $\hat{u}_i$, definidos na Seção 2.3. Mais adiante, exploraremos a relação entre os erros e os resíduos. Para interpretar β_0 e β_1 em uma aplicação particular, (2.47) é mais instrutiva, mas (2.48) também é necessária para algumas derivações estatísticas.

A relação (2.48) pode ser colocada em um gráfico para um registro particular dos dados, como mostrado na Figura 2.7.

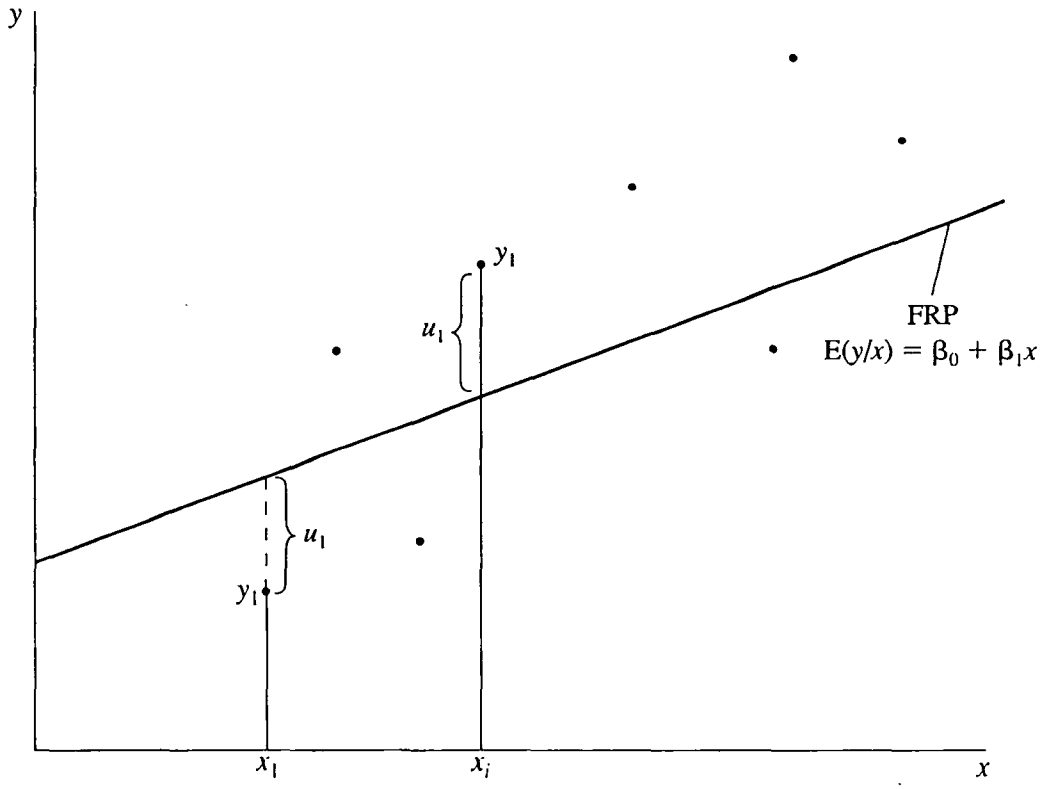
A fim de obter estimadores não-viesados de β_0 e β_1, precisamos impor a hipótese de média condicional zero que discutimos, com algum detalhe, na Seção 2.1. Agora, vamos adicioná-la explicitamente à nossa lista de hipóteses.

HIPÓTESE RLS.3 (MÉDIA CONDICIONAL ZERO)
$E(u|x) = 0$.

Para uma amostra aleatória, essa hipótese implica que $E(u_i|x_i) = 0$, para todo $i = 1,2,..., n$.

Além de restringir a relação entre u e x na população, a hipótese de média condicional zero — juntamente com a hipótese de amostra aleatória — permite uma simplificação técnica conveniente. Em particular, podemos derivar as propriedades estatísticas dos estimadores de MQO como *condicionais* aos valores de x_i em nossa amostra. Tecnicamente, em derivações estatísticas, condicionar aos valores amostrais da variável independente é o mesmo que tratar x_i como *fixo em amostras repetidas*. Esse processo envolve vários passos. Primeiro, escolhemos n valores amostrais para $x_1, x_2, ..., x_n$. (Esses valores podem ser repetidos.) Dados esses valores, obtemos uma amostra de y (efetivamente, obtendo uma amostra aleatória de u_i). Em seguida, obtém-se outra amostra de y, usando novamente os mesmos x_1, x_2, ..., x_n. E assim por diante.

Figura 2.7
Gráfico de $y_i = \beta_0 + \beta_1 x_i + u_i$.



O fixo no cenário de amostras repetidas não é muito realista no contexto não-experimental. Por exemplo, na amostragem de indivíduos do exemplo salários-educação, faz pouco sentido pensar em escolher os valores de *educ* antecipadamente e, em seguida, fazer uma amostra de indivíduos com aqueles níveis particulares de educação formal. A amostragem aleatória, na qual os indivíduos são escolhidos aleatoriamente e seus salários e anos de educação formal são registrados, é um processo representativo de como muitos conjuntos de dados são obtidos para a análise empírica nas ciências sociais. Já que *assumimos* $E(u|x) = 0$, e temos amostragem aleatória, nada se perde nas derivações ao tratar os x_i como não-aleatórios. O perigo é que o fixo na hipótese de amostras repetidas *sempre* implica que u_i e x_i são independentes. Ao decidir quando a análise de regressão simples produzirá estimadores não-viesados, é crucial pensar em termos da hipótese RLS.3.

Visto que concordamos em condicionar as derivações estatísticas aos valores de x_i, precisamos de uma última hipótese para a inexistência de viés.

HIPÓTESE RLS.4 (VARIAÇÃO AMOSTRAL NA VARIÁVEL INDEPENDENTE)
Na amostra, as variáveis independentes x_i, $i = 1,2,...,n$, não são todas iguais a uma mesma constante. Isso exige alguma variação em x na população.

Encontramos a hipótese RLS.4 quando derivamos as fórmulas dos estimadores de MQO; ela é equivalente a $\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$. Das quatro hipóteses feitas, esta é a menos importante, pois ela essen-

cialmente nunca falha em aplicações interessantes. Se a hipótese RLS.4 não se sustentar, não podemos calcular os estimadores de MQO, o que significa que a análise estatística é irrelevante.

Usando o fato de que $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$ (veja o Apêndice A, no site da Thomson), podemos escrever o estimador de inclinação de MQO na equação (2.19) como

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{2.49}$$

Como agora estamos interessados no comportamento de $\hat{\beta}_1$ ao longo de todas as amostras possíveis, $\hat{\beta}_1$ é apropriadamente visto como uma variável aleatória.

Podemos escrever $\hat{\beta}_1$ em termos dos coeficientes populacionais e dos erros ao substituir o lado direito de (2.48) em (2.49). Temos

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\text{SQT}_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\text{SQT}_x}, \tag{2.50}$$

onde definimos a variação total em x_i como $\text{SQT}_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$, a fim de simplificar a notação. (Essa expressão não é exatamente a variância amostral de x_i, pois não a dividimos por $n - 1$.) Usando a álgebra do operador somatório, vamos escrever o numerador de $\hat{\beta}_1$ como

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})\beta_0 + \sum_{i=1}^{n}(x_i - \bar{x})\beta_1 x_i + \sum_{i=1}^{n}(x_i - \bar{x})u_i \\ = \beta_0\sum_{i=1}^{n}(x_i - \bar{x}) + \beta_1\sum_{i=1}^{n}(x_i - \bar{x})x_i + \sum_{i=1}^{n}(x_i - \bar{x})\,u_i.\end{aligned} \tag{2.51}$$

Como mostrado no Apêndice A, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ e $\sum_{i=1}^{n}(x_i - \bar{x})\,x_i = \sum_{i=1}^{n}(x_i - \bar{x})^2 = s_x^2$. Portanto, podemos escrever o numerador de $\hat{\beta}_1$ como $\beta_1\text{SQT}_x + \sum_{i=1}^{n}(x_i - \bar{x})u_i$. Escrevendo isso no denominador resulta

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\text{SQT}_x} = \beta_1 + (1/\text{SQT}_x)\sum_{i=1}^{n} d_i\,u_i, \tag{2.52}$$

onde $d_i = x_i - \bar{x}$. Vemos agora que o estimador $\hat{\beta}_1$ é igual à inclinação populacional β_1 mais um termo que é uma combinação linear dos erros $\{u_1,u_2,...,u_n\}$. Condicionada aos valores de x_i, a aleatoriedade em $\hat{\beta}_1$ deve-se inteiramente aos erros na amostra. O fato de que esses erros sejam, em geral, diferentes de zero é o que faz com que $\hat{\beta}_1$ seja diferente de β_1.

Ao usar a representação em (2.52), podemos provar a primeira importante propriedade estatística do método MQO.

T E O R E M A 2 . 1 (INEXISTÊNCIA DE VIÉS EM MQO)
Usando as hipóteses RLS.1 a RLS.4,

$$\mathrm{E}(\hat{\beta}_0) = \beta_0 \text{ e } \mathrm{E}(\hat{\beta}_1) = \beta_1, \tag{2.53}$$

para quaisquer valores de β_0 e β_1. Em outras palavras, $\hat{\beta}_0$ é não-viesado para β_0, e $\hat{\beta}_1$ é não-viesado para β_1.
PROVA: Nesta prova, os valores esperados estão condicionados aos valores amostrais da variável independente. Visto que SQT_x e d_i são funções somente de x_i, eles são não-aleatórios quando condicionais. Portanto, de (2.52), e mantendo o condicionamento a $\{x_1,x_2,...,x_n\}$ implícito, temos

$$\mathrm{E}(\hat{\beta}_1) = \beta_1 + \mathrm{E}[(1/\mathrm{SQT}_x)\sum_{i=1}^{n} d_i u_i] = \beta_1 + (1/\mathrm{SQT}_x)\sum_{i=1}^{n} \mathrm{E}(d_i u_i)$$

$$= \beta_1 + (1/\mathrm{SQT}_x)\sum_{i=1}^{n} d_i \,\mathrm{E}(u_i) = \beta_1 + (1/\mathrm{SQT}_x)\sum_{i=1}^{n} d_i \cdot 0 = \beta_1,$$

onde usamos o fato de que o valor esperado de cada u_i (condicional a $\{x_1,x_2,...,x_n\}$) é zero sob as hipóteses RLS.2 e RLS.3. Como a inexistência de viés se mantém para qualquer resultado condicionado a $\{x_1,x_2,...,x_n\}$, a inexistência de viés também se mantém sem se condicionar a $\{x_1,x_2,...,x_n\}$.

A prova para $\hat{\beta}_0$ é agora direta. Obtenha a média de (2.48) através de i para obter $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$, e insira essa equação na fórmula de $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \beta_0 + \beta_1 \bar{x} + \bar{u} - \hat{\beta}_1 \bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u}.$$

Então, condicional aos valores de x_i,

$$\mathrm{E}(\hat{\beta}_0) = \beta_0 + \mathrm{E}[(\beta_1 - \hat{\beta}_1)\bar{x}] + \mathrm{E}(\bar{u}) = \beta_0 + \mathrm{E}[(\beta_1 - \hat{\beta}_1)]\bar{x},$$

já que, pelas hipóteses RLS.2 e RLS.3, $\mathrm{E}(\bar{u}) = 0$. No entanto, mostramos que $\mathrm{E}(\hat{\beta}_1) = \beta_1$, o que implica que $\mathrm{E}[(\hat{\beta}_1 - \beta_1)] = 0$. Assim, $\mathrm{E}(\hat{\beta}_0) = \beta_0$. Ambos os argumentos são válidos para quaisquer valores de β_0 e β_1, e assim estabelecemos a inexistência de viés.

Lembre-se de que a inexistência de viés é uma característica das distribuições amostrais de $\hat{\beta}_1$ e $\hat{\beta}_0$, o que não nos diz nada sobre a estimativa que obtemos de uma dada amostra. Esperamos que, se a amostra que obtemos é de algum modo "típica", então nossa estimativa deve estar "próxima" do valor populacional. Infelizmente, é sempre possível obter uma amostra ruim que nos dê uma estimativa pontual distante de β_1, e nós *nunca* saberemos, com certeza, se esse é o caso. Neste ponto, você pode querer revisar o material sobre estimadores não-viesados no Apêndice C (disponível no site da Thomson), especialmente o exercício de simulação da Tabela C.1, o qual ilustra o conceito de inexistência de viés.

Em geral, a inexistência de viés não é válida se qualquer uma das nossas quatro hipóteses não for válida. Isso significa que é importante pensar na veracidade de cada hipótese em uma aplicação particular. Como já discutimos, se a hipótese RLS.4 não for válida, então não seremos capazes de obter as estimativas de MQO. A hipótese RLS.1 requer que y e x sejam linearmente relacionados com uma perturbação adicionada. Certamente, isso pode não ser válido. Mas sabemos também que y e x podem ser escolhidos de forma que possam produzir relações não-lineares interessantes. Estudar com a não validade de (2.47) requer métodos mais avançados, que estão além do escopo deste texto.

Posteriormente, veremos como relaxar a hipótese RLS.2, a hipótese de amostragem aleatória, na análise de séries de tempo. Porém, o que dizer de seu uso na análise de corte transversal? A amostragem aleatória pode não ser válida em um corte transversal quando as amostras não são representativas da população subjacente; de fato, alguns conjuntos de dados são construídos fazendo-se, intencionalmente, amostras de partes diferentes da população. Discutiremos os problemas de amostragem não-aleatória nos Capítulos 9 e 17.

A hipótese na qual devemos nos concentrar agora é a RLS.3. Se RLS.3 se mantém, as estimativas de MQO são não-viesadas. Do mesmo modo, se RLS.3 não se mantém, os estimadores de MQO serão, em geral, viesados. Há maneiras de determinar a direção e o tamanho prováveis do viés, algo que estudaremos no Capítulo 3.

A possibilidade de que x seja correlacionado com u é quase sempre uma preocupação na análise de regressão simples com dados não-experimentais, como indicamos por meio de vários exemplos na Seção 2.1. Usar a regressão simples quando u contém fatores que afetam y e que também estão correlacionados com x pode resultar em *correlação espúria*: isto é, achamos uma relação entre y e x que se deve, em verdade, a outros fatores que afetam y e que também estão correlacionados com x.

EXEMPLO 2.12

(Desempenho em Matemática de Estudantes e o Programa de Merenda Escolar)

Seja *mate10* a percentagem de alunos do primeiro ano do ensino médio aprovados em um exame de matemática. Suponha que desejamos estimar o efeito do programa de merenda escolar financiado pelo governo sobre o desempenho dos alunos. Esperamos que o programa de merenda tenha um efeito *ceteris paribus* positivo sobre o desempenho: todos os outros fatores permanecendo iguais, se um estudante, bastante pobre para ter regularmente refeições, torna-se qualificado para o programa de merenda escolar, seu desempenho deveria melhorar. Seja *prgalm* a percentagem de estudantes que estão aptos para participar do programa de merenda escolar. Portanto, o modelo de regressão simples é

$$mate10 = \beta_0 + \beta_1 prgalm + u, \qquad \textbf{(2.54)}$$

em que u contém características da escola e do estudante que afetam o desempenho escolar total. Usando os dados do arquivo de MEAP93.RAW de 408 escolas de Michigan no ano escolar 1992-1993, obtemos

EXEMPLO 2.12 (continuação)

$$\widehat{maîe10} = 32{,}14 - 0{,}319\, prgalm$$

$$n = 408, R^2 = 0{,}171.$$

Essa equação prevê que se a participação dos estudantes no programa de merenda escolar aumenta em dez pontos percentuais, a percentagem de estudantes que passa no exame de matemática *cai* cerca de 3,2 pontos percentuais. Realmente devemos acreditar que a participação maior no programa de merenda escolar causa, *de fato*, um desempenho pior? Muito provavelmente não. Uma explicação melhor é que o termo erro u na Equação (2.54) está correlacionado com *prgalm*. De fato, u contém fatores como a taxa de pobreza das crianças que freqüentam a escola, que afeta o desempenho dos estudantes e está altamente correlacionada com a qualificação no programa de merenda. Variáveis como qualidade e recursos da escola também estão contidas em u e, provavelmente, estão correlacionados com *prgalm*. É importante lembrar que a estimativa –0,319 é somente para essa amostra particular, mas seu sinal e magnitude nos fazem suspeitar de que u e x sejam correlacionadas, de modo que a regressão linear é viesada.

Além de variáveis omitidas, há outras razões para que x esteja correlacionado com u no modelo de regressão simples. Como essas mesmas questões surgirão na análise de regressão múltipla, postergaremos até lá um tratamento sistemático do problema.

Variâncias dos Estimadores de MQO

Além de saber que a distribuição amostral de $\hat{\beta}_1$ está centrada em torno de β_1 ($\hat{\beta}_1$ é não-viesado), é importante saber o quão distante, em média, podemos esperar que $\hat{\beta}_1$ esteja de β_1. Entre outras coisas, isso nos permite escolher o melhor estimador entre todos os estimadores não-viesados — ou pelo menos entre uma ampla classe deles. A medida de dispersão da distribuição de $\hat{\beta}_1$ (e $\hat{\beta}_0$) com a qual é mais fácil trabalhar é a variância, ou sua raiz quadrada, o desvio-padrão. (Veja o Apêndice C, disponível no site da Thomson, para uma discussão mais detalhada.)

A variância dos estimadores de MQO pode ser calculada sob as hipóteses RLS.1 a RLS.4. Entretanto, as expressões dessas variâncias são complexas. Em vez disso, vamos adicionar uma hipótese que é tradicional na análise de corte transversal. Essa hipótese afirma que a variância do termo não-observável, u, condicionado a x, é constante. Ela é conhecida como a hipótese de **homoscedasticidade** ou de "variância constante".

HIPÓTESE RLS.5 (HOMOSCEDASTICIDADE)

$\text{Var}(u|x) = \sigma^2$.

Devemos enfatizar que a hipótese de homoscedasticidade é completamente distinta da hipótese de média condicional zero, $E(u|x) = 0$. A hipótese RLS.3 compreende o *valor esperado* de u, enquanto a hipótese RLS.5 diz respeito à *variância* de u (ambos condicionados a x). Lembre-se de que nós estabelecemos a inexistência de viés de MQO sem a hipótese RLS.5: a hipótese de homoscedasticidade não desempenha *qualquer* papel para mostrar que $\hat{\beta}_0$ e $\hat{\beta}_1$ são não-viesados. Adicionamos a hipótese

RLS.5 pois ela simplifica os cálculos da variância de $\hat{\beta}_0$ e $\hat{\beta}_1$ e porque ela implica que o método de mínimos quadrados ordinários tenha certas propriedades de eficiência, algo que veremos no Capítulo 3. Se assumíssemos que u e x são *independentes*, a distribuição de u, dado x, não dependeria de x, e assim $E(u|x) = E(u) = 0$ e $Var(u|x) = \sigma^2$. No entanto, algumas vezes independência é, uma hipótese forte demais.

Como $Var(u|x) = E(u^2|x) - [E(u|x)]^2$ e $E(u|x) = 0$, $\sigma^2 = E(u^2|x)$, o que significa que σ^2 também é a esperança *não-condicional* de u^2. Portanto, $\sigma^2 = E(u^2) = Var(u)$, pois $E(u) = 0$. Em outras palavras, σ^2 é a variância *não-condicional* de u, e por isso σ^2 é freqüentemente chamado de **variância do erro** ou variância da perturbação. A raiz quadrada de σ^2, σ, é o desvio-padrão do erro. Um σ grande significa que a distribuição dos fatores não-observáveis que afetam y é mais dispersa.

Freqüentemente, é útil escrever as hipóteses RLS.3 e RLS.5 em termos da média condicional e da variância condicional de y:

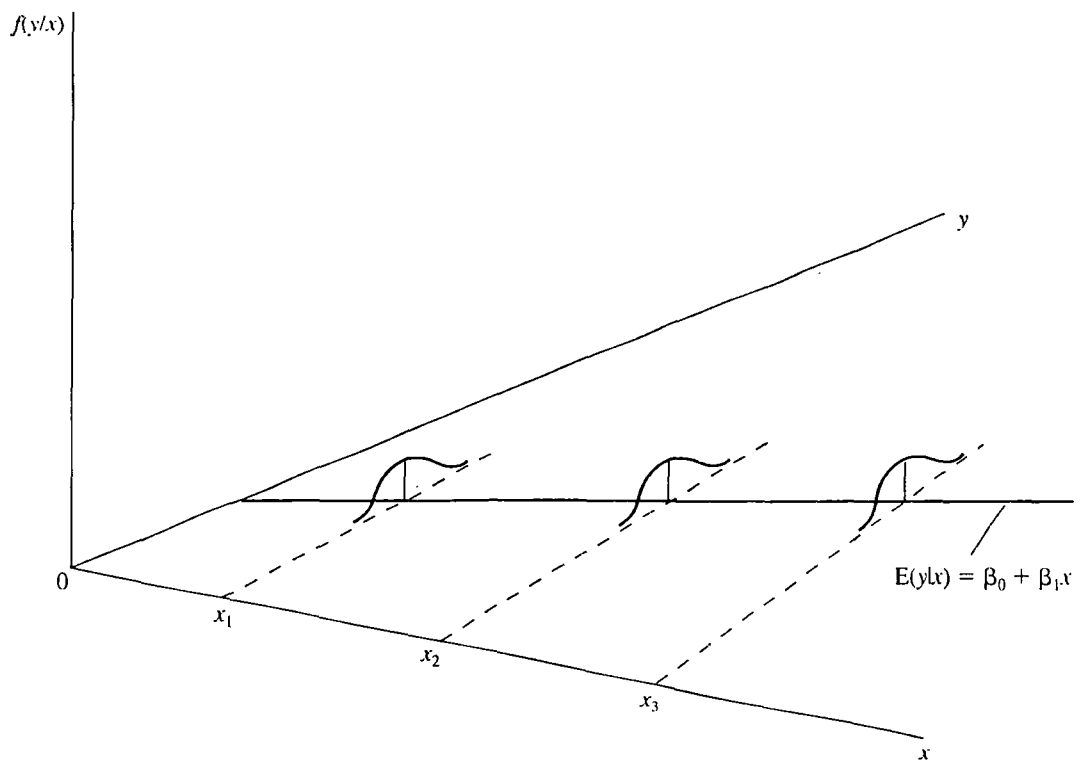
$$E(y|x) = \beta_0 + \beta_1 x. \tag{2.55}$$

$$Var(y|x) = \sigma^2. \tag{2.56}$$

Em outras palavras, a esperança condicional de y, dado x, é linear em x, mas a variância de y, dado x, é constante. Essa situação está ilustrada na Figura 2.8, em que $\beta_0 > 0$ e $\beta_1 > 0$.

Quando $Var(u|x)$ depende de x, diz-se que o termo erro apresenta **heteroscedasticidade** (ou variância não-constante). Como $Var(u|x) = Var(y|x)$, a heteroscedasticidade está presente sempre que $Var(y|x)$ é uma função de x.

Figura 2.8
O modelo de regressão simples sob homoscedasticidade.

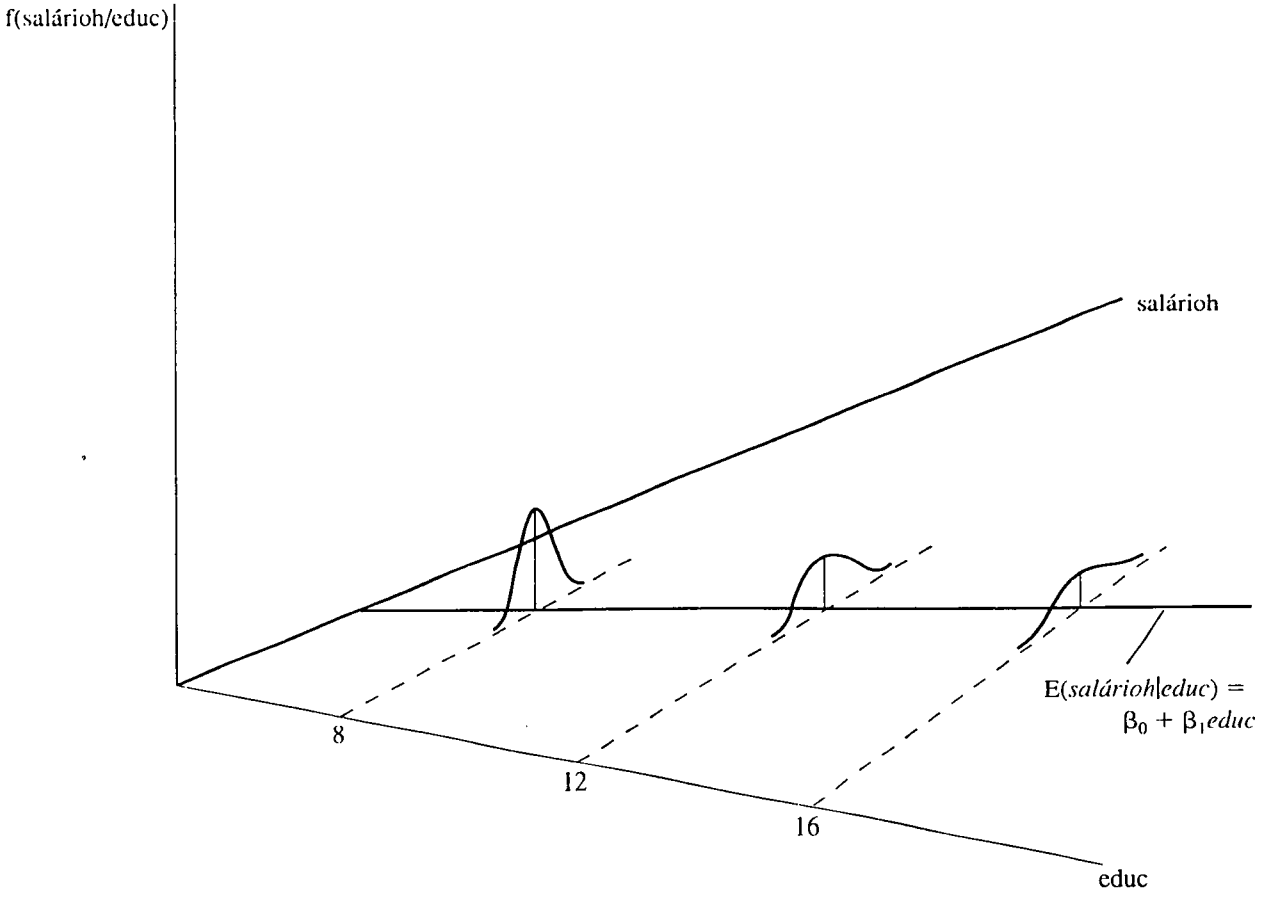


EXEMPLO 2.13

(Heteroscedasticidade em uma Equação de Salários)

A fim de obter um estimador não-viesado do efeito *ceteris paribus* de *educ* sobre *salárioh*, devemos assumir que $E(u|educ) = 0$, e isso implica $E(salárioh|educ) = \beta_0 + \beta_1 educ$. Se também usarmos a hipótese de homoscedasticidade, então $Var(u|educ) = \sigma^2$ não depende do nível de educação formal, que é o mesmo que assumir que $Var(salárioh|educ) = \sigma^2$. Assim, enquanto se deixa o salário-hora médio aumentar com o nível de educação formal – é essa taxa de crescimento que estamos interessados em descrever – assume-se que a *variabilidade* no salário horário em torno de sua média é constante através de todos os níveis de educação formal: isso pode não ser realista. É provável que pessoas com mais tempo de educação formal tenham uma variedade maior de interesses e de oportunidades de trabalho, o que poderia levar a uma variabilidade maior nos níveis de educação formal mais elevados. Pessoas com níveis de educação formal bastante baixos têm muito poucas oportunidades e, freqüentemente, precisam trabalhar recebendo salário mínimo; isso tem o efeito de reduzir a variabilidade salarial nos níveis baixos de educação formal. Essa situação é mostrada na Figura 2.9. Em última análise, se a hipótese RLS.5 se mantém ou não é uma questão empírica. No Capítulo 8 mostraremos como testar a hipótese RLS.5.

Figura 2.9
Var (*salárioh/educ*) crescendo com educ.



Com a hipótese apropriada de homoscedasticidade, estamos prontos para provar o seguinte:

TEOREMA 2.2 (VARIÂNCIAS AMOSTRAIS EM MQO)
Sob as hipóteses RLS.1 a RLS.5,

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sigma^2/\text{SQT}_x, \tag{2.57}$$

e

$$\text{Var}(\hat{\beta}_0) = \frac{\sigma^2 n^{-1}\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{2.58}$$

as quais estão condicionadas aos valores amostrais $\{x_1,...,x_n\}$.

PROVA: Derivaremos a fórmula para $\text{Var}(\hat{\beta}_1)$, deixando a outra derivação como exercício. O ponto de partida é a equação (2.52): $\hat{\beta}_1 = \beta_1 + (1/\text{SQT}_x)\sum_{i=1}^{n} d_i u_i$. Visto que β_1 é exatamente uma constante, condicional aos x_i, SQT_x e $d_i = x_i - \bar{x}$ também são não-aleatórios. Além disso, como os u_i são variáveis aleatórias independentes para todos os i (por amostragem aleatória), a variância da soma é a soma das variâncias. Usando esses fatos, temos

$$\text{Var}(\hat{\beta}_1) = (1/\text{SQT}_x)^2 \text{Var}\left(\sum_{i=1}^{n} d_i u_i\right) = (1/\text{SQT}_x)^2 \left(\sum_{i=1}^{n} d_i^2 \text{Var}(u_i)\right)$$

$$= (1/\text{SQT}_x)^2 \left(\sum_{i=1}^{n} d_i^2 \sigma^2\right) \quad [\text{visto que } \text{Var}(u_i) = \sigma^2 \text{ para todos os } i]$$

$$= \sigma^2 (1/\text{SQT}_x)^2 \left(\sum_{i=1}^{n} d_i^2\right) = \sigma^2 (1/\text{SQT}_x)^2 \text{SQT}_x = \sigma^2/\text{SQT}_x$$

que é o que queríamos provar.

As equações (2.57) e (2.58) são as fórmulas "padrões" para a análise de regressão simples, mas não são válidas na presença de heteroscedasticidade. Isso será importante quando tratarmos dos intervalos de confiança e dos testes de hipóteses na análise de regressão múltipla.

Para a maior parte dos propósitos, estamos interessados em $\text{Var}(\hat{\beta}_1)$. É fácil resumir como essa variância depende da variância do erro, σ^2, e da variação total em $\{x_1, x_2,..., x_n\}$, SQT_x. Primeiro, quanto maior a variância do erro, maior é $\text{Var}(\hat{\beta}_1)$. Isso faz sentido, já que uma variação maior nos fatores

não-observáveis que afetam y faz com que seja mais difícil estimar com precisão β_1. Em contrapartida, é preferível maior variabilidade na variável independente: quando a variabilidade nos x_i aumenta, a variância de $\hat{\beta}_1$ diminui. Isso também tem um sentido intuitivo, visto que quanto mais dispersa for a amostra das variáveis independentes, mais fácil será descrever a relação entre $E(y|x)$ e x. Isto é, será mais fácil estimar β_1. Se há pouca variação nos x_i, pode ser difícil estabelecer com precisão como $E(y|x)$ varia com x. Quando o tamanho da amostra cresce, do mesmo modo cresce a variação total nos x_i. Portanto, um tamanho de amostra maior resulta em uma variância menor de $\hat{\beta}_1$.

Essa análise mostra que, se estamos interessados em $\hat{\beta}_1$, e temos uma escolha, então devemos escolher os x_i tão dispersos quanto possível. Às vezes, isso é possível com dados experimentais, mas raramente temos esse luxo nas ciências sociais: usualmente, temos de pegar os x_i que obtemos via amostragem aleatória. Algumas vezes, temos uma possibilidade de obter amostras maiores, embora isso possa ser dispendioso.

Mostre que, ao estimar β_0, é melhor ter $\bar{x} = 0$. Qual é a $\text{Var}(\hat{\beta}_0)$ nesse caso? [*Sugestão*: para qualquer amostra de números, $\sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$, mantendo-se a igualdade somente se $\bar{x} = 0$.]

Para o propósito de construir intervalos de confiança e derivar estatísticas de testes, precisaremos trabalhar com os desvios padrão de $\hat{\beta}_1$ e $\hat{\beta}_0$, $\text{dp}(\hat{\beta}_1)$ e $\text{dp}(\hat{\beta}_0)$. Lembre-se de que eles são obtidos ao calcular as raízes quadradas das variâncias em (2.57) e (2.58). Particularmente, $\text{dp}(\hat{\beta}_1) = \sigma/\sqrt{\text{SQT}_x}$, em que σ é a raiz quadrada de σ^2 e $\sqrt{\text{SQT}_x}$ é a raiz quadrada de SQT_x.

Estimação da Variância do Erro

As fórmulas em (2.57) e (2.58) permitem-nos isolar os fatores que contribuem para $\text{Var}(\hat{\beta}_1)$ e $\text{Var}(\hat{\beta}_0)$. No entanto, essas fórmulas são desconhecidas, exceto no caso extremamente raro em que σ^2 é conhecido. Não obstante, podemos usar os dados para estimar σ^2, o qual conseqüentemente nos permite estimar $\text{Var}(\hat{\beta}_1)$ e $\text{Var}(\hat{\beta}_0)$.

Este é um bom momento para enfatizar a diferença entre os *erros* (ou perturbações) e os *resíduos*, já que essa discussão é crucial para a construção de um estimador de σ^2. A equação (2.48) mostra como escrever o modelo populacional em termos de uma observação extraída aleatoriamente como $y_i = \beta_0 + \beta_1 x_i + u_i$, em que u_i é o erro da observação i. Podemos expressar também y_i em termos de seu valor estimado e do resíduo, como na equação (2.32): $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$. Comparando essas duas equações, vemos que o erro aparece na equação que contém os parâmetros *populacionais*, β_0 e β_1. Em contrapartida, os resíduos aparecem na equação *estimada* com $\hat{\beta}_0$ e $\hat{\beta}_1$. Os erros nunca são observáveis, enquanto os resíduos são calculados a partir dos dados.

Podemos usar as equações (2.32) e (2.48) para escrever os resíduos como uma função dos erros:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = (\beta_0 - \beta_1 x_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

ou

$$\hat{u}_i = u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) x_i. \qquad \textbf{(2.59)}$$

Embora o valor esperado de $\hat{\beta}_0$ iguale-se a β_0, e similarmente para $\hat{\beta}_1$, $\hat{u}_i$ não é o mesmo que u_i. A diferença entre eles tem, de fato, um *valor esperado* igual a zero.

Agora que entendemos a diferença entre os erros e os resíduos, podemos retornar para estimar σ^2. Primeiro, $\sigma^2 = E(u^2)$, de modo que um "estimador" não-viesado de σ^2 é $n^{-1}\sum_{i=1}^{n} u_i^2$. Infelizmente, esse não é um estimador verdadeiro, pois não observamos os erros u_i. Mas, temos, de fato, as estimativas de u_i, a saber, os resíduos $\hat{u}_i$ de MQO. Se substituímos os erros pelos resíduos de MQO, temos $n^{-1}\sum_{i=1}^{n} \hat{u}_i^2 = \text{SQR}/n$. Esse *é* um estimador verdadeiro, porque ele fornece uma regra computável para qualquer amostra de dados sobre x e y. Uma ligeira desvantagem desse estimador é que ele resulta viesado (embora o viés seja pequeno para n grande). Como é fácil calcular um estimador não-viesado, usamos esse como substituto.

O estimador SQR/n é viesado, essencialmente, porque ele não explica a razão de duas restrições que devem ser satisfeitas pelos resíduos de MQO. Essas restrições são dadas pelas duas condições de primeira ordem de MQO:

$$\sum_{i=1}^{n} \hat{u}_i = 0, \quad \sum_{i=1}^{n} x_i \cdot \hat{u}_i = 0. \tag{2.60}$$

Uma maneira de ver essas restrições é esta: se nós conhecemos $n - 2$ dos resíduos, podemos sempre obter os outros dois resíduos usando as restrições implicadas pelas condições de primeira ordem em (2.60). Assim, há somente $n - 2$ **graus de liberdade** nos resíduos de MQO, em oposição a n graus de liberdade nos erros. Se substituíssemos $\hat{u}_i$ por u_i em (2.60), as restrições não mais se manteriam. O estimador não-viesado de σ^2 que utilizaremos faz um ajustamento dos graus de liberdade:

$$\hat{\sigma}^2 = \frac{1}{(n-2)}\sum_{i=1}^{n} \hat{u}_i^2 = \text{SQR}/(n-2). \tag{2.61}$$

(Esse estimador é, às vezes, denominado por s^2, mas continuaremos a usar a convenção de colocar "chapéus" sobre os estimadores.)

TEOREMA 2.3 (ESTIMAÇÃO NÃO-VIESADA DE σ^2)
Sob as hipóteses RLS.1 a RLS.5,

$$E(\hat{\sigma}^2) = \sigma^2.$$

PROVA: Se construirmos a média da equação (2.59) para todos os i e usarmos o fato de que os resíduos de MQO têm média igual a zero, temos $0 = \bar{u} - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)\bar{x}$; subtraindo essa equação de (2.59), resulta $\hat{u}_i = (u_i - \bar{u}) - (\hat{\beta}_1 - \beta_1)(x_i - \bar{x})$. Portanto, $\hat{u}_i^2 = (u_i - \bar{u})^2 + (\hat{\beta}_1 - \beta_1)^2(x_i - \bar{x})^2 - 2(u_i - \bar{u})(\hat{\beta}_1 - \beta_1)(x_i - \bar{x})$. A soma ao longo de todos os i resulta na equação $\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (u_i - \bar{u})^2 + (\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n} u_i (x_i - \bar{x})$. Agora, o valor esperado do primeiro termo é $(n-1)\sigma^2$, algo que está mostrado no Apêndice C, no site da Thomson.

TEOREMA 2.3 (ESTIMAÇÃO NÃO-VIESADA DE σ^2) (continuação)

O valor esperado do segundo termo é, simplesmente, σ^2, porque $E[(\hat{\beta}_1 - \beta_1)^2] = \text{Var}(\hat{\beta}_1) = \sigma^2/s_x^2$. Finalmente, o terceiro termo pode ser escrito como: $2(\hat{\beta}_1 - \beta_1)^2 s_x^2$; aplicando as esperanças, resulta em $2\sigma^2$. Colocando esses três termos juntos, obtemos $E\left(\sum_{i=1}^{n} \hat{u}_i^2\right) = (n-1)\sigma^2 + \sigma^2 - 2\sigma^2 = (n-2)\sigma^2$, de modo que $E[\text{SQR}/(n-2)] = \sigma^2$.

Se $\hat{\sigma}^2$ for inserido nas fórmulas da variância (2.57) e (2.58), então teremos estimadores não-viesados de $\text{Var}(\hat{\beta}_1)$ e $\text{Var}(\hat{\beta}_0)$. Posteriormente, necessitaremos de estimadores dos desvios padrão de $\hat{\beta}_1$ e $\hat{\beta}_0$, e isso requer estimar σ. O estimador natural de σ é

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \qquad \textbf{(2.62)}$$

e é chamado **erro-padrão da regressão (EPR)**. (Outros nomes para $\hat{\sigma}$ são *erro-padrão da estimativa* e *raiz do erro quadrado médio*, mas não os usaremos.) Ainda que $\hat{\sigma}$ não seja um estimador não-viesado de σ, podemos mostrar que ele é um estimador *consistente* de σ (veja Apêndice C, disponível no site da Thomson), e que ele servirá muito bem para nossos propósitos.

A estimativa $\hat{\sigma}$ é interessante, já que ela é uma estimativa do desvio-padrão dos fatores não observáveis que afetam y; equivalentemente, ela estima o desvio-padrão em y após os efeitos de x terem sido retirados. A maior parte dos programas econométricos informa o valor de $\hat{\sigma}$ juntamente com o R-quadrado, o intercepto, a inclinação e outras estatísticas de MQO (sob um dos vários nomes listados anteriormente). Por enquanto, nosso principal interesse está em usar $\hat{\sigma}$ para estimar os desvios-padrão de $\hat{\beta}_0$ e $\hat{\beta}_1$. Como $\text{dp}(\hat{\beta}_1) = \sigma/s_x$, o estimador natural de $\text{dp}(\hat{\beta}_1)$ é

$$\text{ep}(\hat{\beta}_1) = \hat{\sigma}/s_x = \hat{\sigma}/\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^{1/2},$$

que é chamado de **erro-padrão** de $\hat{\beta}_1$. Observe que $\text{ep}(\hat{\beta}_1)$ é visto como uma variável aleatória quando pensamos em processar o método MQO usando diferentes amostras de y; isso ocorre porque σ^2 varia com diferentes amostras. Para uma dada amostra, $\text{ep}(\hat{\beta}_1)$ é um número, exatamente como $\hat{\beta}_1$ é simplesmente um número quando nós o calculamos a partir de dados conhecidos.

Semelhantemente, $\text{ep}(\hat{\beta}_0)$ é obtido de $\text{dp}(\hat{\beta}_0)$ ao substituir σ por $\hat{\sigma}$. O erro-padrão de qualquer estimativa nos dá uma idéia de qual preciso é o estimador. Os erros-padrão desempenham um papel central em todo este texto; nós os usaremos para construir estatísticas de testes e intervalos de confiança para todos os procedimentos econométricos que cobriremos, a partir do Capítulo 4.

2.6 REGRESSÃO ATRAVÉS DA ORIGEM

Em raros casos, desejamos impor a restrição de que, quando $x = 0$, o valor esperado de y é zero. Há certas relações para as quais isso é razoável. Por exemplo, se a renda (x) for zero, então os gastos com o imposto de renda (y) devem ser zero. Além disso, há problemas quando um modelo que originalmente tem um intercepto diferente de zero é transformado em um modelo sem um intercepto.

Formalmente, nós escolhemos agora um estimador da inclinação, que chamaremos de $\tilde{\beta}_1$, e uma reta da forma

$$\tilde{y} = \tilde{\beta}_1 x, \tag{2.63}$$

em que os sinais gráficos do til sobre $\tilde{\beta}_1$ e $\tilde{y}$ são usados para distinguir esse problema do problema muito mais comum de estimar um intercepto juntamente com uma inclinação. Costuma-se chamar (2.63) de **regressão através da origem**, pois a reta (2.63) passa pelo ponto $x = 0$, $\tilde{y} = 0$. Para obter a estimativa de inclinação em (2.63), nós ainda contamos com o método de mínimos quadrados ordinários, que minimiza, nesse caso, a soma dos resíduos quadrados:

$$\sum_{i=1}^{n} (y_i - \tilde{\beta}_1 x_i)^2. \tag{2.64}$$

Usando cálculo, pode-se mostrar que $\tilde{\beta}_1$ deve resolver a condição de primeira ordem:

$$\sum_{i=1}^{n} x_i (y_i - \tilde{\beta}_1 x_i) = 0. \tag{2.65}$$

Daí, podemos resolver para $\tilde{\beta}_1$:

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}, \tag{2.66}$$

desde que nem todos os x_i sejam zero — um caso que excluímos.

Observe como $\tilde{\beta}_1$ se compara com a estimativa de inclinação quando também estimamos o intercepto (em vez de determiná-lo igual a zero). Essas duas estimativas são as mesmas se, e somente se, $\bar{x} = 0$. [Veja a equação (2.49) para $\hat{\beta}_1$.] Obter uma estimativa de β_1 usando a regressão através da origem não é freqüente em trabalhos aplicados, e por boas razões: se o intercepto $\beta_0 \neq 0$, logo $\tilde{\beta}_1$ é um estimador viesado de β_1. Você será solicitado a provar isso no Problema 2.8.

Neste capítulo introduzimos o modelo de regressão simples e cobrimos suas propriedades básicas. Dada uma amostra aleatória, o método de mínimos quadrados ordinários é usado para estimar os parâmetros de inclinação e de intercepto no modelo populacional. Demonstramos a álgebra da reta de regressão de MQO, incluindo os cálculos dos valores estimados e dos resíduos, e a obtenção das variações previstas na variável dependente a partir de uma dada variação na variável independente. Na Seção 2.4 discutimos duas questões de importância prática: (1) o comportamento das estimativas de MQO quando mudamos as unidades de medida da variável dependente ou da variável independente e (2) o uso do log natural na elaboração de modelos de elasticidade constante e de semi-elasticidade constante.

Na Seção 2.5, mostramos que, sob as quatro hipóteses RLS.1 a RLS.4, os estimadores de MQO são não-viesados. A hipótese fundamental é que o termo erro u tem média zero, dado qualquer valor da variável independente x. Infelizmente, há razões para pensar que isso é falso em muitas aplicações da regressão simples nas ciências sociais, em que os fatores omitidos em u estão freqüentemente correlacionados com x. Quando adicionamos a hipótese de que a variância do erro, dado x, é constante, obtivemos fórmulas simples das variâncias amostrais dos estimadores de MQO. Como vimos, a variância do estimador de inclinação $\hat{\beta}_1$ cresce quando a variância do erro cresce, e ela decresce quando há mais variação amostral na variável independente. Também derivamos um estimador não-viesado para $\sigma^2 = \text{Var}(u)$.

Na Seção 2.6, discutimos brevemente a regressão através da origem, cujo estimador de inclinação é obtido sob a hipótese de que o intercepto é zero. Às vezes, tal regressão é útil, mas ela não aparece com freqüência em trabalhos aplicados.

Temos ainda muito trabalho por fazer. Por exemplo, ainda não sabemos como testar hipóteses sobre os parâmetros populacionais, β_0 e β_1. Assim, embora saibamos que o método MQO é, sob as hipóteses RLS.1 a RLS.4, não-viesado para os parâmetros populacionais, não temos um modo de fazer inferências sobre a população. Outros tópicos, tais como a eficiência de MQO relativa a outros procedimentos possíveis, também foram omitidos.

As questões de intervalos de confiança, testes de hipóteses e eficiência também são centrais para a análise de regressão múltipla. Como a maneira que construímos os intervalos de confiança e as estatísticas de testes é muito similar para a regressão múltipla — e porque a regressão simples é um caso especial da regressão múltipla —, nosso tempo será mais bem gasto se nos movermos para a regressão múltipla, que é muito mais aplicável que a regressão simples. Nosso propósito, no Capítulo 2, foi fazer você pensar nas questões que surgem na análise econométrica dentro de uma estrutura clara e simples.

2.1 Seja *filhos* o número de filhos de uma mulher e *educ* os anos de educação da mulher. Um modelo simples que relaciona a fertilidade a anos de educação é

$$filhos = \beta_0 + \beta_1 educ + u,$$

em que u é um erro não-observável.

(i) Que tipos de fatores estão contidos em u? É provável que eles estejam correlacionados com o nível de educação?

(ii) Uma análise de regressão simples mostrará o efeito *ceteris paribus* da educação sobre a fertilidade? Explique.

2.2 No modelo de regressão linear simples $y = \beta_0 + \beta_1 x + u$, suponha que $E(u) \neq 0$. Fazendo $\alpha_0 = E(u)$, mostre que o modelo pode sempre ser reescrito com a mesma inclinação, mas com um novo intercepto e erro, em que o novo erro tem um valor esperado zero.

2.3 A tabela seguinte contém as variáveis *nmgrad* (nota média em curso superior nos Estados Unidos) e *tac* (nota do teste de avaliação de conhecimentos para ingresso em curso superior nos Estados Unidos) com as notas hipotéticas de oito estudantes de curso superior. A nota *nmgrad* está baseada em uma escala de quatro pontos e foi arredondada para um dígito após o ponto decimal. A nota *tac* baseia-se em uma escala de 36 pontos e foi arredondada para um número inteiro.

Estudante	*nmgrad*	*tac*
1	2,8	21
2	3,4	24
3	3,0	26
4	3,5	27
5	3,6	29
6	3,0	25
7	2,7	25
8	3,7	30

(i) Estime a relação entre *nmgrad* e *tac* usando MQO; isto é, obtenha as estimativas de intercepto e de inclinação da equação

$$\widehat{nmgrad} = \hat{\beta}_0 + \hat{\beta}_1 tac.$$

Comente a direção da relação. O intercepto tem uma interpretação útil aqui? Explique. Qual deveria ser o valor previsto de *nmgrad* se a nota *tac* aumentasse em cinco pontos?

(ii) Calcule os valores estimados e os resíduos de cada observação e verifique que a soma dos resíduos é (aproximadamente) zero.

(iii) Qual é o valor previsto de *nmgrad* quando *tac* = 20?

(iv) Quanto da variação de *nmgrad* dos 8 estudantes é explicada por *tac*? Explique.

2.4 Os dados do arquivo BWGHT.RAW contém dados de nascimentos por mulheres nos Estados Unidos. As duas variáveis de interesse são: a variável dependente, peso dos recém-nascidos em onças* (*pesonas*), e a variável explicativa, número médio de cigarros que a mãe fumou por dia durante a gravidez (*cigs*). A seguinte regressão simples foi estimada usando dados de $n = 1.388$ nascimentos:

$$\widehat{pesonas} = 119{,}77 - 0{,}514\ cigs$$

* NT: 1 onça = 31,10 g.

(i) Qual é o peso de nascimento previsto quando *cigs* = 0? E quando *cigs* = 20 (um maço por dia)? Comente a diferença.

(ii) O modelo de regressão simples necessariamente captura uma relação causal entre o peso de nascimento da criança e os hábitos de fumar da mãe? Explique.

(iii) Para prever um peso de nascimento de 125 onças, qual deveria ser a magnitude de *cigs*? Comente.

(iv) Qual a fração de mulheres na amostra que não fumaram enquanto estiveram grávidas? Isso ajuda a reconciliar sua conclusão da parte (iii)?

2.5 Na função de consumo linear

$$\hat{cons} = \hat{\beta}_0 + \hat{\beta}_1 rend,$$

a *propensão marginal a consumir* PMgC (estimada) é simplesmente a inclinação $\hat{\beta}_1$, enquanto a *propensão média a consumir* PmeC é $\hat{cons}/rend = \hat{\beta}_0/rend + \hat{\beta}_1$. Usando as observações de renda e consumo anuais de 100 famílias (ambas medidas em dólares), obteve-se a seguinte equação:

$$\hat{cons} = -124{,}84 + 0{,}853\, rend$$
$$n = 100,\ R^2 = 0{,}692.$$

(i) Interprete o intercepto dessa equação e comente seu sinal e magnitude.

(ii) Qual é o consumo previsto quando a renda familiar é $ 30.000?

(iii) Com *rend* sobre o eixo de *x*, desenhe um gráfico da PMgC e da PmeC estimadas.

2.6 Usando dados de casas vendidas em 1988 em Andover, Massachusetts [Kiel e McClain (1995)], a equação seguinte relaciona os preços das casas (*preço*) à distância de um incinerador de lixo recentemente construído (*dist*):

$$\widehat{\log(preço)} = 9{,}40 + 0{,}312 \log(dist)$$
$$n = 135,\ R^2 = 0{,}162.$$

(i) Interprete o coeficiente de log(*dist*). O sinal dessa estimativa é o que você esperava?

(ii) Você considera que a regressão simples oferece um estimador não-viesado da elasticidade *ceteris paribus* de *preço* em relação a *dist*? (Pense na decisão da cidade sobre onde colocar o incinerador.)

(iii) Quais outros fatores relativos a casas afetam seu preço? Eles poderiam estar correlacionados com a distância do incinerador?

2.7 Considere a função de poupança

$$poup = \beta_0 + \beta_1 rend + u, \quad u = \sqrt{rend} \cdot e,$$

onde *e* é uma variável aleatória com E(*e*) = 0 e Var(*e*) = σ_e^2. Assuma que *e* é independente de *rend*.

(i) Mostre que $E(u|rend) = 0$, de modo que a hipótese de média condicional zero (hipótese RLS.3) é satisfeita. [*Sugestão*: se e é independente de *rend*, então $E(e|rend) = E(e)$.]

(ii) Mostre que $Var(u|rend) = \sigma_e^2$ *rend*, de modo que a hipótese de homoscedasticidade RLS.5 é violada. Em particular, a variância de *poup* aumenta com *rend*. [*Sugestão*: $Var(e|rend) = Var(e)$, se e e *rend* são independentes.]

(iii) Faça uma discussão que sustente a hipótese de que a variância da poupança aumenta com a renda da família.

2.8 Considere o modelo de regressão simples padrão $y = \beta_0 + \beta_1 x + u$, sob as hipóteses RLS.1 a RLS.4. Os estimadores usuais $\hat{\beta}_0$ e $\hat{\beta}_1$ são não-viesados para seus respectivos parâmetros populacionais. Seja $\tilde{\beta}_1$ o estimador de β_1 obtido ao assumir que o intercepto é zero (veja a Seção 2.6).

(i) Encontre $E(\tilde{\beta}_1)$ em termos de x_i, β_0 e β_1. Verifique que $\tilde{\beta}_1$ é não-viesado para β_1 quando o intercepto populacional (β_0) é zero. Há outros casos em que $\tilde{\beta}_1$ é não-viesado?

(ii) Encontre a variância de $\tilde{\beta}_1$. [*Sugestão*: a variância não depende de β_0.]

(iii) Mostre que $Var(\tilde{\beta}_1) \leq Var(\hat{\beta}_1)$. [*Sugestão*: para qualquer amostra de dados, $\sum_{i=1}^{n} x_i^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$, com a desigualdade estrita preponderando, a não ser que $\bar{x} = 0$.]

(iv) Comente a relação entre viés e variância, ao escolher entre $\hat{\beta}_1$ e $\tilde{\beta}_1$.

2.9 (i) Sejam $\hat{\beta}_0$ e $\hat{\beta}_1$ o intercepto e a inclinação da regressão de y_i sobre x_i, usando n observações. Sejam c_1 e c_2 constantes, com $c_2 \neq 0$. Sejam $\tilde{\beta}_0$ e $\tilde{\beta}_1$ o intercepto e a inclinação da regressão de $c_1 y_i$ sobre $c_2 x_i$. Mostre que $\tilde{\beta}_1 = (c_1/c_2)\hat{\beta}_1$ e $\tilde{\beta}_0 = c_1 \hat{\beta}_0$, verificando as observações sobre as unidades de medida da Seção 2.4. [*Sugestão*: para obter $\tilde{\beta}_1$, insira as transformações de x e y em (2.19). Em seguida, use (2.17) para $\tilde{\beta}_0$, estando seguro de usar as transformações de x e y e a inclinação correta.]

(ii) Agora, sejam $\tilde{\beta}_0$ e $\tilde{\beta}_1$ os parâmetros estimados da regressão de $(c_1 + y_i)$ sobre $(c_2 + x_i)$ (sem qualquer restrição sobre c_1 ou c_2). Mostre que $\tilde{\beta}_1 = \hat{\beta}_1$ e $\tilde{\beta}_0 = \hat{\beta}_0 + c_1 - c_2 \hat{\beta}_1$.

(iii) Agora, sejam $\hat{\beta}_0$ e $\hat{\beta}_1$ as estimativas de MQO da regressão $\log(y_i)$ sobre x_i, para a qual devemos assumir $y_i > 0$ para todo i. Para $c_1 > 0$, sejam $\tilde{\beta}_0$ e $\tilde{\beta}_1$ o intercepto e a inclinação da regressão de $\log(c_1 y_i)$ sobre x_i. Mostre que $\tilde{\beta}_1 = \hat{\beta}_1$ e $\tilde{\beta}_0 = \log(c_1) + \hat{\beta}_0$.

(iv) Agora, assumindo que $x_i > 0$ para todo i, sejam $\tilde{\beta}_0$ e $\tilde{\beta}_1$ o intercepto e a inclinação da regressão de y_i sobre $\log(c_2 x_i)$. Como $\tilde{\beta}_0$ e $\tilde{\beta}_1$ comparam-se com o intercepto e a inclinação da regressão de y_i sobre $\log(x_i)$?

Minimizando a Soma dos Resíduos Quadrados

Mostramos aqui que as estimativas de MQO $\hat{\beta}_0$ e $\hat{\beta}_1$ minimizam a soma dos resíduos quadrados, como afirmado na Seção 2.2. Formalmente, o problema é caracterizar as soluções $\hat{\beta}_0$ e $\hat{\beta}_1$ para o problema de minimização

$$\min_{b_0,b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2,$$

onde b_0 e b_1 são argumentos *dummies* para o problema de otimização; por simplicidade, chame essa função $Q(b_0,b_1)$. De um resultado fundamental do cálculo multivariado (veja Apêndice A, disponível no site de Thomson), uma condição necessária para $\hat{\beta}_0$ e $\hat{\beta}_1$ resolverem o problema de minimização é que as derivadas parciais de $Q(b_0,b_1)$ em relação a b_0 e b_1 devem ser zero quando avaliadas com $\hat{\beta}_0$ e $\hat{\beta}_1$: $\partial Q(\hat{\beta}_0,\hat{\beta}_1)/\partial b_0 = 0$ e $\partial Q(\hat{\beta}_0,\hat{\beta}_1)/\partial b_1 = 0$. Usando a regra da cadeia do cálculo, essas duas equações tornam-se

$$-2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$
$$-2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

Essas duas equações são exatamente (2.14) e (2.15) multiplicadas por $-2n$ e, portanto, são solucionadas pelos mesmos $\hat{\beta}_0$ e $\hat{\beta}_1$.

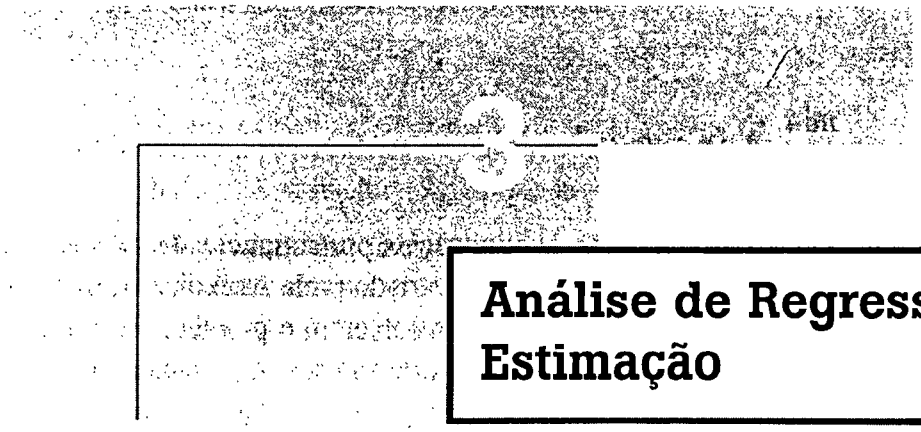
Como sabemos que, realmente, minimizamos a soma dos resíduos quadrados? As condições de primeira ordem são necessárias, mas não são suficientes. Uma maneira de verificar que minimizamos a soma dos resíduos quadrados é escrever, para qualquer b_0 e b_1,

$$\begin{aligned} Q(b_0, b_1) &= \sum_{i=1}^{n} [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i + (\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1) x_i]^2 \\ &= \sum_{i=1}^{n} [\hat{u}_i + (\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1) x_i]^2 \\ &= \sum_{i=1}^{n} \hat{u}_i^2 + n(\hat{\beta}_0 - b_0)^2 + (\hat{\beta}_1 - b_1)^2 \sum_{i=1}^{n} x_i^2 + 2(\hat{\beta}_0 - b_0)(\hat{\beta}_1 - b_1) \sum_{i=1}^{n} x_i, \end{aligned}$$

onde usamos as equações (2.30) e (2.31). A soma dos resíduos quadrados não depende de b_0 e b_1, enquanto a soma dos últimos três termos pode ser escrita como

$$\sum_{i=1}^{n} [(\hat{\beta}_0 - b_0) + (\hat{\beta}_1 - b_1) x_i]^2,$$

como pode ser verificado diretamente por álgebra. Visto que essa expressão é uma soma de termos quadrados, ela é maior que zero. Portanto, seu menor valor ocorre quando $b_0 = \hat{\beta}_0$ e $b_1 = \hat{\beta}_1$.



Análise de Regressão Múltipla: Estimação

o Capítulo 2 aprendemos a usar a análise de regressão simples para explicar uma variável dependente y como função de uma única variável independente x. A desvantagem principal de usar a análise de regressão simples em trabalhos empíricos é o fato de ser muito difícil obter conclusões *ceteris paribus* sobre como x afeta y: a hipótese fundamental, RLS.3 — todos os outros fatores que afetam y são não-correlacionados com x —, é freqüentemente irreal.

A **análise de regressão múltipla** é mais receptiva à análise *ceteris paribus*, pois ela nos permite controlar *explicitamente* muitos outros fatores que, de maneira simultânea, afetam a variável dependente. Isso é importante tanto para testar teorias econômicas quanto para avaliar efeitos da política governamental quando devemos nos basear em dados não-experimentais. Como os modelos de regressão múltipla podem acomodar muitas variáveis explicativas que podem estar correlacionadas, podemos esperar inferir causalidade nos casos em que a análise de regressão simples seria enganosa.

Naturalmente, se adicionarmos ao nosso modelo mais fatores que são úteis para explicar y, então mais da variação de y poderá ser explicada. Assim, a análise de regressão múltipla pode ser usada para construir modelos melhores para prever a variável dependente.

Uma vantagem adicional da análise de regressão múltipla é que ela pode incorporar, completamente, relações de formas funcionais gerais. No modelo de regressão simples, somente a função de uma variável explicativa pode aparecer na equação. Como veremos, o modelo de regressão múltipla permite muito mais flexibilidade.

A Seção 3.1 introduz, formalmente, o modelo de regressão múltipla e, mais adiante, discute as vantagens da regressão múltipla sobre a regressão simples. Na Seção 3.2, demonstramos como estimar os parâmetros do modelo de regressão múltipla por meio do método de mínimos quadrados ordinários. Nas seções 3.3, 3.4 e 3.5 descrevemos várias propriedades dos estimadores de MQO, incluindo a inexistência de viés e a eficiência.

O modelo de regressão múltipla ainda é o veículo mais extensamente usado da análise empírica em economia e em outras ciências sociais. Igualmente, o método de mínimos quadrados ordinários é popularmente usado para estimar os parâmetros do modelo de regressão múltipla.

3.1 FUNCIONALIDADE DA REGRESSÃO MÚLTIPLA

Modelo com Duas Variáveis Independentes

Iniciaremos com alguns exemplos simples para mostrar como a análise de regressão múltipla pode ser usada para resolver problemas que não podem ser resolvidos pela regressão simples.

O primeiro exemplo é uma variação simples da equação do salário introduzida no Capítulo 2 para obter o efeito da educação sobre o salário-hora:

$$salárioh = \beta_0 + \beta_1 educ + \beta_2 exper + u, \quad (3.1)$$

em que *exper* representa anos de experiência no mercado de trabalho. Assim, *salárioh* é determinado por duas variáveis explicativas ou independentes, educação e experiência, e por outros fatores não-observados, contidos em *u*. Basicamente, ainda estamos interessados no efeito de *educ* sobre *salárioh*, mantendo fixos todos os outros fatores que afetam *salárioh*; isto é, estamos interessados no parâmetro β_1.

Comparada com uma análise de regressão simples que relaciona *salárioh* a *educ*, a equação (3.1) remove, efetivamente, *exper* do termo erro e a coloca explicitamente na equação. Como *exper* aparece na equação, seu coeficiente, β_2, mede o efeito *ceteris paribus* de *exper* sobre *salárioh*, que também é de algum interesse.

Não surpreendentemente, assim como na regressão simples, teremos de fazer hipóteses sobre como *u*, em (3.1), está relacionado às variáveis independentes, *educ* e *exper*. Entretanto, como veremos na Seção 3.2, há uma coisa da qual podemos estar seguros: visto que (3.1) contém a experiência de modo explícito, seremos capazes de mensurar o efeito da educação sobre o salário horário, mantendo a experiência fixa. Na análise de regressão simples — que coloca *exper* no termo erro —, teríamos de assumir que experiência é não-correlacionada com educação, uma hipótese tênue.

Como segundo exemplo, considere o problema de explicar o efeito do gasto público por estudante (*gasto*) sobre a nota média padronizada (*notmed*) do ensino médio. Suponha que a nota média dependa do gasto público, da renda familiar média (*rendfam*) e de outros fatores não-observáveis:

$$notmed = \beta_0 + \beta_1 gasto + \beta_2 rendfam + u. \quad (3.2)$$

Para o propósito de análise da política governamental, o coeficiente de interesse é β_1, o efeito *ceteris paribus* de gasto sobre *notmed*. Ao incluir *rendfam* explicitamente no modelo, somos capazes de controlar seu efeito sobre *notmed*. Isso é provavelmente importante, pois a renda familiar média tende a estar correlacionada com o gasto público por estudante: os níveis de gasto público são, freqüentemente, determinados tanto por impostos locais sobre a propriedade como sobre a renda. Na análise de regressão simples, *rendfam* estaria incluída no termo erro, que estaria provavelmente correlacionado com *gasto*, fazendo com que o estimador de β_1 de MQO fosse viesado no modelo de duas variáveis.

Nesses dois exemplos similares, mostramos que outros fatores observáveis, além da variável de interesse primordial [*educ* na equação (3.1) e *gasto* na equação (3.2)] podem ser incluídos em um modelo de regressão. Em geral, podemos escrever um modelo com duas variáveis independentes como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \quad (3.3)$$

em que β_0 é o intercepto, β_1 mede a variação em *y* com relação a x_1, mantendo fixos outros fatores, e β_2 mede a variação em *y* com relação a x_2, mantendo outros fatores fixos.

A análise de regressão múltipla também é útil para generalizar relações funcionais entre variáveis. Como exemplo, suponha que o consumo da família (*cons*) é uma função quadrática da renda familiar (*rend*):

$$cons = \beta_0 + \beta_1 rend + \beta_2 rend^2 + u, \qquad (3.4)$$

em que u contém outros fatores que afetam o consumo. Nesse modelo, o consumo depende somente de um fator observado, a renda; desse modo, pareceria que ele pode ser tratado dentro do arcabouço da regressão simples. No entanto, esse modelo está fora do padrão da regressão simples, porque ele contém duas funções da renda, *rend* e $rend^2$ (e, portanto, três parâmetros: β_0, β_1 e β_2). Portanto, a função consumo é facilmente escrita como modelo de regressão com duas variáveis, e fazendo $x_1 = rend$ e $x_2 = rend^2$.

Mecanicamente, não há *nenhuma* diferença em usar o método de mínimos quadrados ordinários (introduzido na Seção 3.2) para estimar equações tão diferentes como (3.1) e (3.4). Cada uma dessas equações pode ser escrita como (3.3), que é tudo o que importa para os cálculos. Há, entretanto, uma diferença importante em como *interpretar* os parâmetros. Na equação (3.1), β_1 é o efeito *coeteris paribus* de *educ* sobre *salárioh*. O parâmetro β_1 não tem tal interpretação em (3.4). Em outras palavras, não faz sentido medir o efeito de *rend* sobre *cons* mantendo, ao mesmo tempo, $rend^2$ fixo, porque se *rend* varia, então $rend^2$ deve variar! Em vez disso, a variação no consumo com respeito à variação na renda — a propensão marginal a consumir — é aproximada por

$$\frac{\Delta cons}{\Delta rend} \approx \beta_1 + 2\beta_2 rend$$

Para os cálculos necessários quanto a derivação dessa equação, veja o Apêndice A, disponível na página do livro, no site www.thomsonlearning.com.br. Em outras palavras, o efeito marginal da renda sobre o consumo depende tanto de β_2 como de β_1 e do nível de renda. Esse exemplo mostra que, em qualquer aplicação particular, as definições das variáveis independentes são cruciais. Mas, para o desenvolvimento teórico da regressão múltipla, podemos ser vagos com relação a tais detalhes. No Capítulo 6 estudaremos exemplos como esse de forma mais completa.

No modelo com duas variáveis independentes, a hipótese fundamental sobre como u está relacionado a x_1 e x_2 é

$$E(u|x_1, x_2) = 0. \qquad (3.5)$$

A interpretação da condição (3.5) é similar à interpretação da hipótese RLS.3 da análise de regressão simples. Ela significa que, para qualquer valor de x_1 e x_2 na população, o fator não-observável médio é igual a zero. Assim como na regressão simples, a parte importante da hipótese é que o valor esperado de u é o mesmo para todas as combinações de x_1 e x_2; dizer que esse valor comum é zero está longe de ser uma hipótese, desde que o intercepto β_0 esteja incluído no modelo (veja a Seção 2.1).

Como podemos interpretar a hipótese de média condicional zero no exemplo anterior? Na equação (3.1), a hipótese é $E(u|educ,exper) = 0$. Isso implica que outros fatores que afetam *salárioh* não estão, em média, relacionados a *educ* e *exper*. Portanto, se entendermos que aptidão inata é parte de u, então precisaremos que os níveis médios de aptidão sejam os mesmos em todas as combinações de educação e experiência na população que trabalha. Isso pode ou não ser verdadeiro, mas, como veremos na Seção 3.3, essa é a questão que precisamos fazer a fim de determinar se o método de mínimos quadrados ordinários produz estimadores não-viesados.

O exemplo que mede o desempenho dos estudantes [equação (3.2)] é similar à equação do salário. A hipótese de média condicional zero é $E(u|gasto,rendfam) = 0$, o que significa que os outros fato-

res que afetam as notas — características das escolas e dos estudantes — são, em média, não-relacionados aos gastos públicos por estudante e à renda familiar média.

Um modelo simples para explicar as taxas de homicídio nas cidades (*taxahom*) em termos da probabilidade de condenação (*prcond*) e da duração média da sentença (*sentmed*) é

$$taxhom = \beta_0 + \beta_1 prcond + \beta_2 sentmed + u.$$

Que fatores estão contidos em u? Você entende ser provável que a hipótese (3.5) se mantenha?

Quando aplicada à função quadrática do consumo em (3.4), a hipótese da média condicional zero tem uma interpretação ligeiramente diferente. A equação (3.5), escrita literalmente, é $E(u|rend,rend^2) = 0$. Como $rend^2$ é conhecido quando se conhece *rend*, é redundante incluir $rend^2$ na esperança: $E(u|rend,rend^2) = 0$ é o mesmo que $E(u|rend) = 0$. Não há problema em colocar $rend^2$ junto com *rend* na esperança quando expressamos a hipótese, mas $E(u|rend) = 0$ é mais conciso.

Modelo com *k* Variáveis Independentes

Como estamos no contexto da regressão múltipla, não há necessidade de ficarmos com duas variáveis independentes. A análise de regressão múltipla permite que muitos fatores observados afetem *y*. No exemplo do salário, poderíamos também incluir semanas de treinamento de trabalho, anos de permanência com o empregador atual, medidas de aptidão e mesmo variáveis demográficas, como o número de irmãos ou a educação da mãe. No exemplo do gasto público por estudante, poderiam ser incluídos variáveis adicionais que medissem a qualidade dos professores e o tamanho das escolas.

O **modelo de regressão linear múltipla** geral (também chamado modelo de regressão múltipla) pode ser escrito, na população, como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k + u, \qquad \textbf{(3.6)}$$

onde β_0 é o **intercepto**, β_1 é o parâmetro associado a x_1, β_2 é o parâmetro associado a x_2, e assim por diante. Como há *k* variáveis independentes e um intercepto, a equação (3.6) contém $k + 1$ parâmetros (desconhecidos) populacionais. Para simplificar, algumas vezes vamos nos referir aos outros parâmetros que não o intercepto como **parâmetros de inclinação**, ainda que, literalmente, nem sempre eles tenham esse significado. [Veja a equação (3.4), em que nenhum dos parâmetros, β_1 e β_2, é, por si mesmo, uma inclinação, mas juntos determinam a inclinação da relação entre o consumo e a renda.]

A terminologia da regressão múltipla é similar àquela da regressão simples e é apresentada na Tabela 3.1. Exatamente como na regressão simples, a variável *u* é o **termo erro** ou **perturbação**. Ele contém outros fatores, além de $x_1, x_2, ..., x_k$, que afetam *y*. Não importa quantas variáveis explicativas incluímos em nosso modelo, pois sempre haverá fatores que não podemos incluir, e eles estão contidos, coletivamente, em *u*.

Tabela 3.1

Terminologia para a Regressão Múltipla

y	$x_1, x_2, ..., x_k$
Variável Dependente	Variáveis Independentes
Variável Explicada	Variáveis Explicativas
Variável de Resposta	Variáveis de Controle
Variável Prevista	Variáveis Previsoras
Regressando	Regressores

Ao aplicar o modelo de regressão múltipla geral, devemos saber como interpretar os parâmetros. Agora e nos capítulos subseqüentes vamos adquirir bastante prática, mas é útil, neste ponto, relembrarmos algumas coisas que já sabemos. Suponha que os salários (*salário*) dos diretores executivos estejam relacionados às vendas das empresas (vendas) e à permanência dos diretores executivos nas empresas (*permceo*) pela equação

$$\log(salário) = \beta_0 + \beta_1 \log(vendas) + \beta_2 permceo + \beta_3 permceo^2 + u. \qquad (3.7)$$

Essa equação enquadra-se no modelo de regressão múltipla (com $k = 3$) ao definirmos $y = \log(salário)$, $x_1 = \log(vendas)$, $x_2 = permceo$ e $x_3 = permceo^2$. Como sabemos do Capítulo 2, o parâmetro β_1 é a elasticidade (*ceteris paribus*) de *salário* em relação a *vendas*. Se $\beta_3 = 0$, então $100\,\beta_2$ é, aproximadamente, o aumento percentual *ceteris paribus* em *salário* quando *permceo* aumenta em um ano. Quando $\beta_3 \neq 0$, o efeito de *permceo* sobre *salário* é mais complicado. Postergaremos até o Capítulo 6 um tratamento detalhado de modelos com termos quadráticos.

A equação (3.7) fornece um lembrete importante sobre a análise de regressão múltipla. O termo "linear" na expressão "modelo de regressão linear múltipla" significa que a equação (3.6) é linear nos *parâmetros*, β_j. A equação (3.7) é um exemplo de modelo de regressão múltipla que, ao mesmo tempo, é linear nos β_j e é uma relação não-linear entre *salário* e as variáveis *vendas* e *permceo*. Muitas aplicações da regressão múltipla envolvem relações não-lineares entre as variáveis subjacentes.

É fácil expressar a hipótese essencial para o modelo de regressão múltipla geral em termos de uma esperança condicional:

$$E(u|x_1, x_2, ..., x_k) = 0. \qquad (3.8)$$

No mínimo, a equação (3.8) requer que todos os fatores no termo erro não-observado sejam não-correlacionados com as variáveis explicativas. Ela também significa que consideramos corretamente a relação funcional entre as variáveis explicada e as explicativas. Qualquer problema que faça com que u seja correlacionado com qualquer variável independente fazem com que (3.8) não seja válida. Na Seção 3.3 mostraremos que a hipótese (3.8) implica que o método MQO é não-viesado e derivaremos o viés que surge quando uma variável-chave for omitida da equação. Nos capítulos 15 e 16 estudaremos outras razões que podem fazer com que (3.8) não seja válida e mostraremos o que pode ser feito nesses casos.

3.2 MECÂNICA E INTERPRETAÇÃO DOS MÍNIMOS QUADRADOS ORDINÁRIOS

Vamos resumir, agora, algumas características computacionais e algébricas do método de mínimos quadrados ordinários, quando ele se aplica a um conjunto particular de dados. Discutiremos também como interpretar a equação estimada.

Obtenção das Estimativas de MQO

Vamos considerar, primeiramente, a estimação do modelo com duas variáveis independentes. A equação de MQO estimada é escrita de forma similar ao caso da regressão simples:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2, \tag{3.9}$$

onde $\hat{\beta}_0$ é a estimativa de β_0, $\hat{\beta}_1$ é a estimativa de β_1, e $\hat{\beta}_2$ é a estimativa de β_2. Porém, como obtemos $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$? O método de **mínimos quadrados ordinários** escolhe as estimativas que minimizam a soma dos resíduos quadrados. Isto é, dadas n observações de y, x_1 e x_2, $\{(x_{i1},x_{i2},y_i)$: $i = 1, 2, ..., n\}$, as estimativas $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$ são escolhidas, simultaneamente, para fazer com que a expressão

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2})^2 \tag{3.10}$$

tenha o menor tamanho possível.

A fim de entender o que o meetodo MQO está fazendo, é importante dominar o significado da indexação das variáveis independentes em (3.10). As variáveis independentes têm, aqui, dois subscritos: i seguido de 1 ou 2. O subscrito i refere-se ao número da observação. Assim, a soma em (3.10) contempla todas as observações de $i = 1$ a n. O segundo índice é simplesmente um método para distinguir as diferentes variáveis independentes. No exemplo que relaciona *salárioh* a *educ* e *exper*, $x_{i1} = educ_i$ é a educação formal da pessoa i na amostra, e $x_{i2} = exper_i$ é a experiência da pessoa i. A soma dos resíduos quadrados na equação (3.10) é $\sum_{i=1}^{n} (salário_i - \hat{\beta}_0 - \hat{\beta}_1 educ_i - \hat{\beta}_2 exper_i)^2$. No que vem a seguir, o subscrito i é reservado para indexar o número da observação. Se escrevermos x_{ij}, então isso significa a i-ésima observação da j-ésima variável independente. (Alguns autores preferem mudar a ordem do número da observação e do número da variável, de modo que x_{1i} é a observação i da variável um. Mas isso é apenas um problema de gosto notacional.)

No caso geral com k variáveis independentes, procuramos estimar $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$ na equação

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k. \tag{3.11}$$

As $k + 1$ estimativas de MQO delas foram escolhidas para minimizar a soma dos resíduos quadrados:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - ... - \hat{\beta}_k x_{ik})^2. \tag{3.12}$$

Esse problema de minimização pode ser resolvido usando cálculo multivariado (veja o Apêndice 3A, disponível no site da Thomson. Isso leva a $k + 1$ equações lineares com $k + 1$ estimadores $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ desconhecidos:

$$\begin{aligned}
&\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - ... - \hat{\beta}_k x_{ik}) = 0 \\
&\sum_{i=1}^{n} x_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - ... - \hat{\beta}_k x_{ik}) = 0 \\
&\sum_{i=1}^{n} x_{i2} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - ... - \hat{\beta}_k x_{ik}) = 0 \qquad \textbf{(3.13)} \\
&\vdots \\
&\sum_{i=1}^{n} x_{ik} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - ... - \hat{\beta}_k x_{ik}) = 0.
\end{aligned}$$

Essas equações são,, chamadas freqüentemente de **condições de primeira ordem** de MQO. Assim, como no modelo de regressão simples da Seção 2.2, as condições de primeira ordem de MQO podem ser obtidas pelo método dos momentos: sob a hipótese (3.8), $E(u) = 0$ e $E(x_j u) = 0$, onde $j = 1, 2, ..., k$. As equações em (3.13) são contrapartidas amostrais desses momentos da população, embora tenhamos omitido a divisão pelo tamanho da amostra n.

Mesmo para n e k de tamanhos moderados, resolver as equações em (3.13) fazendo os cálculos manualmente é tedioso. Não obstante, computadores modernos que processam programas padrões de estatística e econometria podem resolver essas equações com n e k grandes muito rapidamente.

Há somente um pequeno aviso: devemos assumir que as equações em (3.13) podem ser resolvidas *unicamente* para os $\hat{\beta}_j$. Por enquanto, assumimos apenas isso, como é usualmente o caso em modelos bem definidos. Na Seção 3.3 formulamos a hipótese necessária para a existência de estimativas de MQO únicas (veja a hipótese RLM.4).

Como na análise de regressão simples, a equação (3.11) é chamada **reta de regressão de MQO** ou a **função de regressão amostral (FRA)**. Chamaremos $\hat{\beta}_0$ a **estimativa de intercepto de MQO** e $\hat{\beta}_1, ..., \hat{\beta}_k$ de **estimativas de inclinação de MQO** (correspondentes às variáveis independentes x_1, $x_2, ..., x_k$).

A fim de indicar que uma regressão de MQO foi computada, escreveremos a equação (3.11) com y e $x_1, ..., x_k$ substituídos pelos seus nomes de variável (tais como *salárioh*, *educ* e *exper*), ou diremos que "rodamos uma regressão de MQO de y sobre $x_1, x_2, ..., x_k$ ou que regredimos y sobre $x_1, x_2, ..., x_k$". Essas expressões são modos de dizer que o método de mínimos quadrados ordinários foi usado para obter a equação de MQO (3.11). A não ser que afirmemos explicitamente, sempre estimaremos um intercepto juntamente com as inclinações.

Interpretação da Equação de Regressão de MQO

Mais importante que os detalhes subjacentes à computação dos $\hat{\beta}_j$ é a *interpretação* da equação estimada. Iniciaremos com o caso de duas variáveis independentes:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2. \qquad \textbf{(3.14)}$$

O intercepto $\hat{\beta}_0$ na equação (3.14) é o valor previsto de y quando $x_1 = 0$ e $x_2 = 0$. Às vezes, colocar x_1 e x_2 iguais a zero é um cenário interessante; em outros casos, isso não fará sentido. Não obstante, para obter uma previsão de y a partir da reta de regressão de MQO, o intercepto sempre é necessário, como (3.14) deixa claro.

As estimativas $\hat{\beta}_1$ e $\hat{\beta}_2$ têm interpretações de **efeito parcial**, ou ***ceteris paribus***. Da equação (3.14), temos

$$\Delta\hat{y} = \hat{\beta}_1\Delta x_1 + \hat{\beta}_2\Delta x_2,$$

de modo que podemos obter a variação prevista em y dadas as variações em x_1 e x_2. (Observe que o intercepto não tem nada a ver com as variações em y.) Em particular, quando x_2 é mantido fixo, de modo que $\Delta x_2 = 0$, então

$$\Delta\hat{y} = \hat{\beta}_1\Delta x_1,$$

mantendo x_2 fixo. O ponto fundamental é que, ao incluir x_2 no nosso modelo, obtemos um coeficiente de x_1 com uma interpretação *ceteris paribus*. Essa é a razão de a análise de regressão múltipla ser tão útil. Semelhantemente,

$$\Delta\hat{y} = \hat{\beta}_2\Delta x_2,$$

mantendo x_1 fixo.

EXEMPLO 3.1

(Determinantes da Nota Média em Curso Superior nos Estados Unidos)

As variáveis do arquivo GPA1.RAW incluem a nota média em um curso superior (*nmgrad*), a nota média do ensino médio (*nmem*) e a nota do teste de avaliação de conhecimentos para ingresso em curso superior (*tac*) para uma amostra de 141 estudantes de uma grande universidade dos Estados Unidos; tanto *nmgrad* como *nmem* estão baseados em uma escala de quatro pontos. Obtemos a seguinte reta de regressão de MQO para estimar *nmgrad* a partir de *nmem* e *tac*:

$$\widehat{nmgrad} = 1{,}29 + 0{,}453\ nmem + 0{,}0094\ tac. \qquad \textbf{(3.15)}$$

Como interpretamos essa equação? Primeiro, o intercepto de 1,29 é o valor previsto de *nmgrad* se tanto *nmem* como *tac* forem iguais a zero. Como ninguém que freqüenta um curso superior teve nota média no ensino médio igual a zero ou uma nota no teste de ingresso no curso superior igual a zero, o intercepto nessa equação não é, por si mesmo, significativo.

As estimativas mais interessantes são os coeficientes de inclinação de *nmem* e *tac*. Como esperado, há uma relação parcial positiva entre *nmgrad* e *nmem*: mantendo *tac* fixo, um ponto adicional em *nmem* está associado a 0,453 de um ponto em *nmgrad*, ou quase meio ponto. Em outras palavras, se escolhermos dois estudantes, A e B, e esses estudantes tiverem a mesma nota *tac*, mas *nmem* do estudante A é um ponto maior que a *nmem* do estudante B, prevemos que o estudante A tem *nmgrad*

EXEMPLO 3.1 (continuação)

0,453 maior que *nmgrad* do estudante B. (Isso não diz nada sobre quaisquer duas pessoas reais, mas é a nossa melhor previsão.)

O sinal *tac* implica que, mantendo *nmem* fixo, uma variação de 10 pontos na nota em *tac* — uma variação muito grande, visto que a nota média na amostra é de cerca de 24, com um desvio-padrão menor que três — afeta *nmgrad* em menos de um décimo de um ponto. Esse é um efeito pequeno e sugere que, uma vez considerada o *nmem*, a nota do *tac* não é um forte previsor de *nmgrad*. (Naturalmente, há muitos outros fatores que contribuem para *nmgrad*, mas aqui estamos enfatizamos as estatísticas disponíveis de estudantes do ensino médio.) Posteriormente, após discutirmos a inferência estatística, mostraremos que o coeficiente de *tac* não é somente pequeno na prática, mas ele também é estatisticamente não significante.

Se colocarmos o foco na análise de regressão simples relacionando somente *nmgrad* e *tac*, obtemos

$$nmgrad = 2{,}40 + 0{,}0271\ tac;$$

assim, o coeficiente *tac* é quase três vezes maior que a estimativa em (3.15). No entanto, essa equação *não* nos permite comparar duas pessoas com o mesmo *nmem*; ela corresponde a um experimento diferente. Mais adiante, falaremos mais sobre as diferenças entre as regressões múltipla e simples.

O caso com mais de duas variáveis independentes é similar. A reta de regressão de MQO é

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \tag{3.16}$$

Escrita em termos de variações,

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \ldots + \hat{\beta}_k \Delta x_k. \tag{3.17}$$

O coeficiente de x_1 mede a variação em $\hat{y}$ devido a um aumento de uma unidade em x_1, mantendo todas as outras variáveis independentes fixas. Isto é,

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1, \tag{3.18}$$

mantendo x_2, x_3, ..., x_k fixos. Assim, controlamos as variáveis x_2, x_3, ..., x_k ao estimar o efeito de x_1 sobre y. Os outros coeficientes têm uma interpretação similar.

O exemplo seguinte contém três variáveis independentes.

EXEMPLO 3.2

(Equação do Salário Horário)

Usando as 526 observações de trabalhadores do arquivo WAGE1.RAW, incluímos *educ* (anos de educação formal), *exper* (anos de experiência no mercado de trabalho) e *perm* (anos com o empregador atual) na equação que explica log(*salárioh*). A equação estimada é

EXEMPLO 3.2 (continuação)

$$\widehat{\log(salário)} = 0{,}284 + 0{,}092\, educ + 0{,}0041\, exper + 0{,}022\, perm. \qquad \textbf{(3.19)}$$

Como no caso da regressão simples, os coeficientes têm uma interpretação de percentagem. A única diferença é que eles também têm uma interpretação *ceteris paribus*. O coeficiente 0,092 significa que, mantendo *exper* e *perm* fixos, um ano a mais de educação formal aumenta o valor esperado de log(*salárioh*) em 0,092, o que se traduz em um aumento aproximado de 9,2% [100(0,092)] em *salárioh*. Alternativamente, se considerarmos duas pessoas com os mesmos níveis de experiência e permanência no trabalho, o coeficiente de *educ* é a diferença proporcional no salário horário previsto quando seus níveis de educação diferem em um ano. Essa medida de retorno da educação mantém fixos ao menos dois importantes fatores de produtividade; saber se ela é uma boa estimativa do retorno *ceteris paribus* de mais um ano de educação formal requer que estudemos as propriedades estatísticas de MQO (veja a Seção 3.3).

Sobre o Significado de "Manter Outros Fatores Fixos" na Regressão Múltipla

Como a interpretação de efeito parcial dos coeficientes de inclinação na análise de regressão múltipla pode causar alguma confusão, assim, vamos tentar impedir o surgimento desse problema agora.

No Exemplo 3.1, observamos que o coeficiente *tac* mede a diferença prevista em *nmgrad*, mantendo *nmem* fixo. O poder da análise de regressão múltipla é que ela proporciona uma interpretação *ceteris paribus* mesmo que os dados *não* sejam coletados de uma maneira *ceteris paribus*. Ao dar ao coeficiente de *tac* uma interpretação de efeito parcial, pode parecer que, realmente, saímos a campo e extraímos amostras compostas de pessoas com a mesma *nmem* e, possivelmente, com diferentes notas do *tac*. Isso não é verdade. Os dados são uma amostra aleatória de uma universidade grande: não há restrições colocadas sobre os valores amostrais de *nmem* ou *tac* na obtenção dos dados. De fato, raramente temos o luxo de manter certas variáveis fixas na obtenção de nossa amostra. *Se* pudéssemos coletar uma amostra de indivíduos com a mesma *nmem*, então poderíamos realizar uma análise de regressão simples relacionando *nmgrad* a *tac*. A regressão múltipla nos permite, efetivamente, simular essa situação sem restringir os valores de quaisquer variáveis independentes.

O poder que a análise de regressão múltipla tem é que ela nos permite fazer, em ambientes não-experimentais, o que os cientistas naturais são capazes de fazer em um ambiente controlado de laboratório: manter outros fatores fixos.

Variação de mais de uma Variável Independente Simultaneamente

Às vezes, queremos variar mais que uma variável independente ao mesmo tempo para encontrar o efeito resultante sobre a variável dependente. Isso é facilmente feito usando a equação (3.17). Por exemplo, na equação (3.19), podemos obter o efeito estimado sobre *salárioh* quando um indivíduo permanece na mesma empresa por mais um ano: ambos *exper* (experiência geral da força de trabalho) e *perm* aumentam em um ano. O efeito total (mantendo *educ* fixo) é

$$\Delta\widehat{\log(salário)} = 0{,}0041\, \Delta exper + 0{,}022\, \Delta perm = 0{,}0041 + 0{,}022 = 0{,}0261,$$

ou cerca de 2,6%. Como *exper* e *perm* aumentam, cada um, em um ano, apenas somamos os coeficientes de *exper* e *perm* e multiplicamos por 100 para converter o efeito em uma percentagem.

Valores Estimados e Resíduos de MQO

Após obter a reta de regressão de MQO (3.11), podemos obter um *valor ajustado* ou *predito* para cada observação. Para a observação i, o valor ajustado é simplesmente

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}, \qquad \textbf{(3.20)}$$

que é exatamente o valor previsto obtido inserindo os valores das variáveis independentes da observação i na equação (3.11). Ao obter os valores estimados, não devemos esquecer do intercepto; de outro modo, a resposta pode ser muito equivocada. Como exemplo, se em (3.15) $nmem_i = 3{,}5$ e $tac_i = 24$, $nmgrad_i = 1{,}29 + 0{,}453(3{,}5) + 0{,}0094(24) = 3{,}101$ (arredondado em três casas após o decimal).

Normalmente, para qualquer observação i, o valor real y_i não se iguala ao valor previsto, $\hat{y}_i$. O método de MQO minimiza o erro quadrado *médio* de previsão, que não diz nada sobre o erro de previsão de qualquer observação específica. O **resíduo** da observação i é definido exatamente como no caso da regressão simples,

$$\hat{u}_i = y_i - \hat{y}_i. \qquad \textbf{(3.21)}$$

Há um resíduo para cada observação. Se $\hat{u}_i > 0$, então $\hat{y}_i$ está abaixo de y_i, o que significa que, para essa observação, y_i é subestimado. Se $\hat{u}_i < 0$, então $y_i < \hat{y}_i$, e y_i é superestimado.

Os valores estimados de MQO e os resíduos têm algumas propriedades importantes que são extensões imediatas do caso da variável única:

1. A média amostral dos resíduos é zero.
2. A covariância amostral entre cada variável independente e os resíduos de MQO é zero. Conseqüentemente, a covariância amostral entre os valores estimados de MQO e os resíduos de MQO é zero.
3. O ponto $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k, \bar{y})$ está sempre sobre a reta de regressão de MQO: $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \ldots + \hat{\beta}_k \bar{x}_k$.

As duas primeiras propriedades são conseqüências imediatas do conjunto de equações usadas para obter as estimativas de MQO. A primeira equação em (3.13) diz que a soma dos resíduos é zero. As equações restantes são da forma $\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0$, implicando que cada variável independente tem covariância amostral zero com $\hat{u}_i$. A Propriedade 3 decorre imediatamente da Propriedade 1.

No Exemplo 3.1, a reta estimada de MQO que explica *nmgrad* em termos de *nmem* é

$$nmgrad = 1{,}29 + 0{,}453\ nmem + 0{,}0094\ tac.$$

Se *nmem* médio é de cerca de 3,4 e a nota média do *tac* está em torno de 24,2, qual é o *nmgrad* médio na amostra?

Interpretação de "Parcialidade" da Regressão Múltipla

Ao aplicar o método MQO, não precisamos saber as fórmulas explícitas dos $\hat{\beta}_j$ que solucionam o sistema de equações em (3.13). Entretanto, para certas derivações, precisamos de fórmulas explícitas dos $\hat{\beta}_j$. Essas fórmulas também ajudam a esclarecer o funcionamento de MQO.

Considere, uma vez mais, o caso com $k = 2$ variáveis independentes, em que $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. Para uma idéia mais concreta, vamos enfatizar $\hat{\beta}_1$. Um modo de expressar $\hat{\beta}_1$ é

$$\hat{\beta}_1 = \left(\sum_{i=1}^{n} \hat{r}_{i1} y_i\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right), \qquad \textbf{(3.22)}$$

onde os $\hat{r}_{i1}$ são os resíduos de MQO de uma regressão simples de x_1 sobre x_2, usando a amostra à mão. Regredimos nossa primeira variável independente, x_1, sobre nossa segunda variável independente, x_2, e, em seguida, obtemos os resíduos (y não desempenha nenhum papel aqui). A equação (3.2) mostra que podemos, portanto, fazer uma regressão simples de y sobre $\hat{r}_1$ para obter $\hat{\beta}_1$. (Observe que os resíduos $\hat{r}_{i1}$ têm uma média amostral zero, e assim $\hat{\beta}_1$ é a estimativa de inclinação usual da regressão simples.)

A representação da equação (3.22) dá outra demonstração da interpretação do efeito parcial de $\hat{\beta}_1$. Os resíduos $\hat{r}_{i1}$ são a parte de x_{i1} que é não-correlacionada com x_{i2}. Outro modo de dizer isso é que $\hat{r}_{i1}$ é x_{i1} após o efeito de x_{i2} ter sido *isolado*, ou *deduzido*. Assim, $\hat{\beta}_1$ mede a relação amostral entre y e x_1 após x_2 ter sido imparcializado.

Na análise de regressão simples, não há parcialização de outras variáveis, porque outras variáveis não estão incluídas na regressão. O Problema 3.17 o conduzirá, passo a passo, pelo processo de parcialização usando os dados de salários do Exemplo 3.2. Para propósitos práticos, o importante é que $\hat{\beta}_1$, na equação $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, mede a variação em y, dado um aumento de uma unidade em x_1, mantendo x_2 fixo.

No modelo geral com k variáveis explicativas, $\hat{\beta}_1$ pode ainda ser escrito como na equação (3.22), mas os resíduos $\hat{r}_{i1}$ vêm da regressão de x_1 sobre x_2, ..., x_k. Assim, $\hat{\beta}_1$ mede o efeito de x_1 sobre y após x_2, ..., x_k terem sido isolados ou deduzidos.

Comparação das Estimativas das Regressões Simples e Múltipla

Há dois casos especiais em que tanto a regressão simples de y sobre x_1 como a regressão de y sobre x_1 e x_2 produzirão a mesma estimativa de MQO de x_1. Para maior precisão, escreva a regressão simples de y sobre x_1 como $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$, e escreva a regressão múltipla como $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. Sabemos que o coeficiente da regressão simples $\tilde{\beta}_1$ não se iguala, geralmente, ao coeficiente da regressão múltipla $\hat{\beta}_1$. Acontece que há uma relação simples entre $\tilde{\beta}_1$ e $\hat{\beta}_1$, que permite comparações interessantes entre as regressões simples e múltipla:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1,$$

em que $\tilde{\delta}_1$ é o coeficiente de inclinação da regressão simples de x_{i2} sobre x_{i1}, $i = 1, ..., n$. Essa equação mostra como $\tilde{\beta}_1$ difere do efeito parcial de x_1 sobre $\hat{y}$. O termo que pode causar confusão é o efeito parcial de x_2 sobre $\hat{y}$ multiplicado pela inclinação da regressão amostral de x_2 sobre x_1. (Veja a Seção 3A.4, no apêndice do capítulo, disponível no site da Thomson, para uma verificação mais geral.)

A relação entre $\tilde{\beta}_1$ e $\hat{\beta}_1$ também mostra que há dois casos distintos em que eles são iguais:

1. O efeito parcial de x_2 sobre $\hat{y}$ é zero na amostra, isto é, $\hat{\beta}_2 = 0$.
2. x_1 e x_2 são não-correlacionados na amostra, isto é, $\tilde{\delta}_1 = 0$.

Ainda que as estimativas das regressões múltipla e simples quase nunca sejam idênticas, podemos usar a fórmula anterior para caracterizar o motivo pelo qual elas deveriam ser muito diferentes ou bastante similares. Por exemplo, se $\hat{\beta}_2$ é pequeno, deveríamos esperar que as estimativas das regressões múltipla e simples de β_1 sejam semelhantes. No Exemplo 3.1, a correlação amostral entre *nmem* e *tac* é de cerca de 0,346, o que não é uma correlação trivial. Porém, o coeficiente de *tac* é razoavelmente pequeno. Não é surpreendente descobrir que a regressão de *nmgrad* sobre *nmem* produz uma estimativa de inclinação de 0,482, não é muito diferente da estimativa de 0,453 em (3.15).

EXEMPLO 3.3

(Participação nos Planos de Pensão 401(k))

Vamos usar os dados do arquivo em 401K.RAW para estimar o efeito de uma taxa de contribuição para um plano (*taxcont*) sobre a taxa de participação (*taxap*) dos trabalhadores nos planos de pensão de contribuição definida existentes nos Estados Unidos. A taxa de contribuição é a quantidade com a qual a firma contribui para um fundo de trabalhadores, para cada dólar de contribuição do trabalhador (até certo limite); assim, *taxcont* = 0,75 significa que a firma contribui com 75 centavos de dólar para cada dólar contribuído pelo trabalhador. A taxa de participação é a percentagem de trabalhadores habilitados a ter uma conta no plano de pensão. A variável *idade* é a idade do plano de pensão. Há 1.534 planos no banco de dados, a *taxap* média é 87,36, a *taxcont* média é 0,732 e a *idade* média é 13,2.

Regredindo *taxap* sobre *taxcont* e *idade* resulta na equação

$$\widehat{taxap} = 80{,}12 + 5{,}52\, taxcont + 0{,}243\, idade. \qquad \textbf{(3.23)}$$

Assim, *taxcont* e *idade* têm os efeitos esperados. O que aconteceria se não controlássemos a variável *idade*? O efeito estimado de *idade* não é trivial, e portanto poderíamos esperar uma variação grande no efeito estimado de *taxcont* se *idade* fosse excluída da regressão. Entretanto, a regressão simples de *taxap* sobre *taxcont* produz $\widehat{taxap} = 83{,}08 + 5{,}86\, taxcont$. A estimativa de regressão simples do efeito de *taxcont* sobre *taxap* é, claramente, diferente da estimativa de regressão múltipla, mas a diferença não é muito grande. (A estimativa da regressão simples é somente cerca de 6,2% maior que a estimativa da regressão múltipla.) Isso pode ser explicado pelo fato de a correlação amostral entre *taxcont* e *idade* ser somente de 0,12.

No caso com k variáveis independentes, a regressão simples de y sobre x_1 e a regressão múltipla de y sobre $x_1, x_2, \ldots, x_k$ produzem uma estimativa idêntica de x_1 somente se: (1) os coeficientes de MQO de x_2 até x_k forem todos zero *ou* (2) x_1 for não-correlacionado com *cada um dos* $x_2, \ldots, x_k$. Na prática, nenhuma dessas possibilidades é muito provável. Porém, se os coeficientes de x_2 até x_k forem pequenos, ou as correlações amostrais entre x_1 e as outras variáveis independentes forem pouco substanciais, então as estimativas das regressões simples e múltiplas do efeito de x_1 sobre y podem ser similares.

Grau de Ajuste

Assim como na regressão simples, podemos definir a **Soma dos Quadrados Total (SQT)**, a **Soma dos Quadrados Explicada (SQE)** e a **Soma dos Quadrados dos Resíduos** ou **Soma dos Resíduos Quadrados (SQR)** como

$$\text{SQT} \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3.24}$$

$$\text{SQE} \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{3.25}$$

$$\text{SQR} \equiv \sum_{i=1}^{n} \hat{u}_i^2. \tag{3.26}$$

Usando o mesmo argumento utilizado no caso da regressão simples, podemos mostrar que

$$\text{SQT} = \text{SQE} + \text{SQR}. \tag{3.27}$$

Em outras palavras, a variação total em $\{y_i\}$ é a soma das variações totais em $\{\hat{y}_i\}$ e em $\{\hat{u}_i\}$.

Assumindo que a variação total em y não é zero — como é o caso, a não ser que y_i seja constante na amostra —, podemos dividir (3.27) por SQT para obter

$$\text{SQR/SQT} + \text{SQE/SQT} = 1.$$

Exatamente como no caso da regressão simples, o R-quadrado é definido como

$$R^2 \equiv \text{SQE/SQT} = 1 - \text{SQR/SQT}, \tag{3.28}$$

e é interpretado como a proporção da variação amostral em y_i que é explicada pela reta de regressão de MQO. Por definição, R^2 é um número entre zero e um.

Pode-se também mostrar que R^2 é igual ao quadrado do coeficiente de correlação entre os valores reais y_i real e os valores estimados $\hat{y}_i$ ajustado. Isto é,

$$R^2 = \frac{\left(\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2\right)}. \tag{3.29}$$

[Inserimos a média de $\hat{y}_i$ em (3.29) por coerência com a fórmula do coeficiente de correlação; sabemos que essa média é igual a $\bar{y}$, porque a média amostral dos resíduos é zero e $y_i = \hat{y}_i + \hat{u}_i$.]

Um fato importante sobre R^2 é que ele nunca diminui, e geralmente aumenta, quando outra variável independente é adicionada à regressão. Esse fato algébrico ocorre por definição, pois a

soma dos resíduos quadrados nunca aumenta quando regressores adicionais são acrescentados ao modelo.

O fato de que R^2 nunca diminui quando *qualquer* variável for adicionada a uma regressão faz dele um instrumento fraco para decidir se uma variável ou diversas variáveis deveriam ser adicionadas ao modelo. O fator que deve determinar se uma variável explicativa pertence a um modelo é se a variável explicativa tem, na *população*, um efeito parcial sobre y diferente de zero. No Capítulo 4, quando cobrirmos a inferência estatística, mostraremos como testar essa hipótese. Veremos também que, quando usado apropriadamente, R^2 permite-nos *testar* um grupo de variáveis com a finalidade de ver se ele é importante para explicar y. Por enquanto, usaremos R^2 como uma medida do grau de ajuste para um dado modelo.

EXEMPLO 3.4

(Determinantes de *nmgrad*)

Da regressão de *nmgrad* que fizemos anteriormente, a equação com R^2 é

$$nmgrad = 1{,}29 + 0{,}453\, nmem + 0{,}0094\, tac$$
$$n = 141, R^2 = 0{,}176.$$

Isso significa que *nmen* e *tac* explicam, juntos, cerca de 17,6% da variação em *nmgrad* da amostra de estudantes. Isso pode não parecer uma percentagem alta, mas devemos nos lembrar de que há muitos outros fatores — incluindo formação da família, personalidade, qualidade da educação do ensino médio, afinidade com o curso escolhido — que contribuem para o desempenho dos estudantes. Se *nmem* e *tac* explicassem quase toda a variação em *nmgrad*, então o desempenho no curso superior seria predeterminado pelo desempenho no ensino médio!

EXEMPLO 3.5

(Explicando os Registros de Prisões)

O arquivo CRIME1.RAW contém dados de prisões durante o ano de 1986 e outras informações sobre 2.725 homens nascidos em 1960 ou 1961 na Califórnia. Cada homem na amostra foi preso pelo menos uma vez antes de 1986. A variável *npre86* é o número de vezes que determinado homem foi preso em 1986: ela é zero para muitos homens da amostra (72,29%), e varia de 0 a 12. (A percentagem de homens presos uma vez em 1986 foi de 20,51%.) A variável *pcond* é a proporção (não a percentagem) de prisões anteriores a 1986 que levaram à condenação, *sentmed* é a duração média da sentença cumprida por condenação prévia (zero para muitas pessoas), *ptemp86* são os meses passados na prisão em 1986 e *empr86* é o número de trimestres durante o qual determinado homem ficou empregado em 1986 (de zero a quatro).

Um modelo linear explicando as prisões é

$$npre86 = \beta_0 + \beta_1 pcond + \beta_2 sentmed + \beta_3 ptemp86 + \beta_4 empr86 + u,$$

em que *pcond* é uma variável (*proxy*) da probabilidade de um homem ser condenado por um crime, e *sentmed* é uma medida do rigor esperado da pena, em caso de condenação. A variável *ptemp86* captura o efeito de confinamento do crime: se um indivíduo está na prisão, ele não pode ser preso por um crime fora da prisão. As oportunidades no mercado de trabalho são capturadas grosseiramente por *empr86*.

Primeiro, estimamos o modelo sem a variável *sentmed*. Obtemos

EXEMPLO 3.5 (continuação)

$$\widehat{npre86} = 0{,}712 - 0{,}150\, pcond - 0{,}034\, ptemp86 - 0{,}104\, empr86$$
$$n = 2.725, R^2 = 0{,}0413.$$

Essa equação diz que, como um grupo, as três variáveis *pcond*, *ptemp86* e *empr86* explicam cerca de 4,1% da variação em *npre86*.

Cada um dos coeficientes de inclinação de MQO tem o sinal esperado. Um aumento na proporção de condenações diminui o número previsto de prisões. Se aumentarmos *pcond* em 0,50 (um aumento grande na probabilidade de condenação), então, mantendo os outros fatores fixos, $\Delta\widehat{npre86} = -0{,}150(0{,}50) = -0{,}075$. Isso pode parecer pouco usual, porque uma prisão não pode ser uma fração. No entanto, podemos usar esse valor para obter a variação prevista das prisões esperadas de um grande grupo de homens. Por exemplo, entre cem homens, a queda esperada de prisões quando *pcond* aumenta em 0,50 é −7,5.

Semelhantemente, um período de prisão mais longo leva a um número previsto menor de prisões. De fato, se *ptemp86* aumenta de 0 para 12, as prisões previstas para um determinado homem diminuem em 0,034(12) = 0,408. Um trimestre a mais no qual o emprego legal é informado diminui as prisões esperadas em 0,104, o que significaria 10,4 prisões entre cem homens.

Se *sentmed* for adicionado ao modelo, sabemos que R^2 aumentará. A equação estimada é

$$\widehat{npre86} = 0{,}707 - 0{,}151\, pcond + 0{,}0074\, sentmed - 0{,}037\, ptemp86 - 0{,}103\, empr86$$
$$n = 2.725, R^2 = 0{,}0422.$$

Assim, ao adicionar a variável sentença média, R^2 aumenta de 0,0413 para 0,0422, um efeito praticamente insignificante. O sinal do coeficiente de *sentmed* também é inesperado: ele diz que uma duração mais longa da sentença média aumenta a atividade criminal.

O Exemplo 3.5 merece uma palavra final de cautela. O fato de as quatro variáveis explicativas incluídas na segunda regressão explicarem somente 4,2% da variação em *npre86* não necessariamente significa que a equação é inútil. Ainda que, coletivamente, essas variáveis não expliquem muito da variação nas prisões, é possível que as estimativas de MQO sejam estimativas confiáveis dos efeitos *ceteris paribus* de cada variável independente sobre *npre86*. Como veremos, se esse for o caso, isso não depende, diretamente, do tamanho do R^2. Em geral, um R^2 baixo indica que é difícil prever resultados individuais sobre y com muita precisão, algo que estudaremos com mais detalhes no Capítulo 6. No exemplo da prisão, o R^2 pequeno reflete algo sobre o qual suspeitamos nas ciências sociais: geralmente, é muito difícil prever o comportamento individual.

Regressão através da Origem

Algumas vezes, uma teoria econômica, ou o senso comum, sugere que β_0 deveria ser zero, e por isso devemos mencionar, brevemente, a estimação de MQO quando o intercepto é zero. Especificamente, vamos agora buscar uma equação da forma

$$\tilde{y} = \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \ldots + \tilde{\beta}_k x_k, \qquad \textbf{(3.30)}$$

em que o símbolo gráfico "~" sobre as estimativas é utilizado para distingui-las das estimativas de MQO obtidas juntamente com o intercepto [como em (3.11)]. Em (3.30), quando $x_1 = 0, x_2 = 0, \ldots, x_k = 0$, o

valor previsto é zero. Nesse caso, diz-se que $\tilde{\beta}_1, ..., \tilde{\beta}_k$ são as estimativas de MQO da regressão de y sobre $x_1, x_2, ..., x_k$ *através da origem.*

As estimativas de MQO em (3.30), como sempre, minimizam a soma dos resíduos quadrados, mas com o intercepto igualado a zero. Você deve estar prevenido de que as propriedades de MQO que derivamos anteriormente não se mantêm mais para a regressão através da origem. Em particular, os resíduos de MQO não têm mais uma média amostral zero. Além disso, se R^2 for definido como 1 − SQR/SQT, em que SQT está dado em (3.24) e SQR é agora $\sum_{i=1}^{n} (y_i - \tilde{\beta}_1 x_{i1} - ... - \tilde{\beta}_k x_{ik})^2$, então R^2 pode ser, de fato, negativo. Isso significa que a média amostral, $\bar{y}$, "explica" mais da variação em y_i do que as variáveis explicativas.

Devemos incluir um intercepto na regressão ou devemos concluir que as variáveis explicativas explicam fracamente y. A fim de sempre ter um R-quadrado não-negativo, alguns economistas preferem calcular R^2 como o quadrado do coeficiente de correlação entre os valores reais e estimados de y, como em (3.29). (Nesse caso, o valor estimado médio deve ser calculado diretamente, já que ele não é mais igual a $\bar{y}$.) Entretanto, não há um conjunto de regras sobre como calcular o R-quadrado para a regressão através da origem.

Uma desvantagem séria com a regressão através da origem é que, se o intercepto β_0 for diferente de zero no modelo populacional, então os estimadores dos parâmetros de inclinação serão viesados. O viés pode ser severo em alguns casos. O custo de estimar um intercepto quando β_0 é realmente zero é que as variâncias dos estimadores de inclinação de MQO são maiores.

3.3 O VALOR ESPERADO DOS ESTIMADORES DE MQO

Vamos nos voltar, agora, para as propriedades estatísticas do método de MQO, para estimar os parâmetros de um modelo da população subjacente. Nesta seção, derivamos o valor esperado dos estimadores de MQO. Em particular, formulamos e discutimos quatro hipóteses, que são extensões diretas das hipóteses do modelo de regressão simples, sob as quais os estimadores de MQO são estimadores não-viesados dos parâmetros da população. Também obtemos explicitamente o viés em MQO, quando uma variável importante for omitida da regressão.

Você deve lembrar que propriedades estatísticas não têm nada a ver com uma amostra particular, mas sim, mais precisamente, com a propriedade dos estimadores quando a amostragem aleatória é feita repetidamente. Assim, as seções 3.3, 3.4 e 3.5 são um pouco abstratas. Apesar de darmos exemplos de derivação do viés de modelos específicos, não é significativo falar sobre as propriedades estatísticas de um conjunto de estimativas de uma única amostra.

A primeira hipótese que vamos fazer define, simplesmente, o modelo de regressão linear múltipla (RLM).

HIPÓTESE RLM.1 (LINEAR NOS PARÂMETROS)

O modelo na população pode ser escrito como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u, \qquad \textbf{(3.31)}$$

em que $\beta_0, \beta_1, ..., \beta_k$ são os parâmetros desconhecidos (constantes) de interesse, e u é um erro aleatório não-observável ou um termo de perturbação aleatória.

A equação (3.31) especifica, formalmente, o **modelo populacional**, algumas vezes chamado **modelo verdadeiro**, para considerar a possibilidade de podermos estimar um modelo diferente de (3.31). A característica fundamental é que o modelo é linear nos parâmetros β_0, β_1, ..., β_k. Como sabemos, (3.31) é bastante flexível, pois y e as variáveis independentes podem ser funções arbitrárias de variáveis subjacentes de interesse, como os logaritmos naturais e os quadrados [veja, por exemplo, a equação (3.7)].

HIPÓTESE RLM.2 (AMOSTRAGEM ALEATÓRIA)
Temos uma amostra aleatória de n observações, $\{(x_{i1},x_{i2},...,x_{ik},y_i): i = 1,2,...,n\}$, do modelo populacional descrito por (3.31).

Às vezes, precisamos escrever a equação de uma observação particular i: para uma observação extraída aleatoriamente da população, temos

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i. \qquad \textbf{(3.32)}$$

Lembre-se de que i refere-se à observação, enquanto o segundo subscrito em x é o número da variável. Por exemplo, podemos escrever uma equação do salário de diretores executivos para um diretor executivo específico particular i como

$$\log(salário) = \beta_0 + \beta_1 \log(vendas_i) + \beta_2 permceo_i + \beta_3 perceo_i^2 + u_i. \qquad \textbf{(3.33)}$$

O termo u_i contém os fatores não-observáveis para o diretor executivo i que afetam seu salário. Nas aplicações, é usualmente mais fácil escrever o modelo na forma populacional, como em (3.31). Ela é menos desordenada e enfatiza que estamos interessados em estimar a relação populacional.

À luz do modelo (3.31), os estimadores de MQO $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, ..., $\hat{\beta}_k$ da regressão de y sobre x_1, ..., x_k são agora considerados estimadores de β_0, β_1, ..., β_k. Vimos, na Seção 3.2, que MQO escolhe as estimativas de uma amostra particular, de modo que os resíduos sejam, em média, iguais a zero e a correlação amostral entre cada variável independente e os resíduos seja zero. Para que MQO seja não-viesado, é preciso que a versão *populacional* dessa condição seja verdadeira.

HIPÓTESE RLM.3 (MÉDIA CONDICIONAL ZERO)
O erro u tem um valor esperado igual a zero, dados quaisquer valores das variáveis independentes. Em outras palavras,

$$E(u|x_1,x_2, ...,x_k) = 0. \qquad \textbf{(3.34)}$$

Uma maneira como a hipótese RLM.3 pode ser violada é quando a relação funcional entre as variáveis explicadas e explicativas está mal-especificada na equação (3.31): por exemplo, se esquece-

mos de incluir o termo quadrático $rend^2$ na função consumo $cons = \beta_0 + \beta_1 rend + \beta_2 rend^2 + u$ quando estimamos o modelo. Outra forma funcional mal-especificada ocorre quando usamos o nível da variável e, de fato, é o log da variável que aparece no modelo populacional, ou vice-versa. Por exemplo, se o modelo verdadeiro tiver log(*salário*) como variável dependente, mas usarmos *salário* como variável dependente em nossa análise de regressão, então os estimadores serão viesados. Intuitivamente, isso deveria ser muito claro. No Capítulo 9 discutiremos maneiras de detectar a má especificação da forma funcional.

Omitir um fator importante que está correlacionado com qualquer uma das variáveis $x_1, x_2, ..., x_k$ faz com que a hipótese RLM.3 também não se sustente. Com a análise de regressão múltipla, somos capazes de incluir muitos fatores entre as variáveis explicativas e, por isso, variáveis omitidas são menos prováveis de serem um problema na análise de regressão múltipla do que na análise de regressão simples. Não obstante, em qualquer aplicação, há sempre muitos fatores que, devido à limitação de dados ou à ignorância deles, não somos capazes de incluir. Se acharmos que esses fatores devem ser controlados e que eles estão correlacionados com uma ou mais variáveis independentes, então a hipótese RLM.3 será violada. Posteriormente, derivaremos esse viés.

Há outros modos pelos quais u pode estar correlacionado com uma variável explicativa. No Capítulo 15 discutiremos o problema do erro de medida em uma variável explicativa. No Capítulo 16 cobriremos o problema, conceitualmente mais difícil, em que uma ou mais variáveis explicativas é determinada conjuntamente com y. Vamos postergar nosso estudo desses problemas até que tenhamos um domínio firme da análise de regressão múltipla sob um conjunto ideal de hipóteses.

Quando a hipótese RLM.3 se mantém, dizemos freqüentemente que temos **variáveis explicativas exógenas**. Se x_j for correlacionado com u por alguma razão, então se diz que x_j é uma **variável explicativa endógena**. Os termos "exógena" e "endógena" originaram-se da análise de equações simultâneas (veja Capítulo 16), mas o significado do termo "variável explicativa endógena" evoluiu e passou a incluir qualquer caso em que uma variável explicativa pode estar correlacionada com o termo erro.

A última hipótese de que precisamos para mostrar que MQO é não-viesado assegura que os estimadores de MQO são, realmente, bem definidos. Para a regressão simples, precisamos assumir que a única variável independente não era constante na amostra. A hipótese correspondente para a análise de regressão múltipla é mais complicada.

HIPÓTESE RLM.4 (COLINEARIDADE NÃO PERFEITA)

Na amostra (e, portanto, na população), nenhuma das variáveis independentes é constante, e não há relações *lineares exatas* entre as variáveis independentes.

A hipótese de colinearidade não perfeita somente diz respeito às variáveis independentes. Estudantes de econometria iniciantes tendem a confundir as hipóteses RLM.4 e RML.3, de modo que enfatizamos aqui que RLM.4 não diz *nada* sobre a relação entre u e as variáveis explicativas.

A hipótese RLM.4 é mais complicada que sua contrapartida na regressão simples, pois agora devemos examinar as relações entre todas as variáveis independentes. Se uma variável independente em (3.31) é uma combinação linear exata de outras variáveis independentes, dizemos que o modelo sofre de **colinearidade perfeita**, e ele não pode ser estimado por MQO.

É importante observar que a hipótese RLM.4 permite, *de fato*, que as variáveis independentes sejam correlacionadas; elas apenas não podem ser correlacionadas *perfeitamente*. Se não permitíssemos qualquer correlação entre as variáveis independentes, então a regressão múltipla não seria muito útil para a análise econométrica. Por exemplo, no modelo que relaciona notas de estudantes aos gastos com educação e à renda familiar,

$$notmed = \beta_0 + \beta_1 gasto + \beta_2 rendfam + u,$$

esperamos, com certeza, que *gasto* e *rendfam* sejam correlacionados: distritos escolares com rendas familiares médias altas tendem a gastar mais em educação por estudante. De fato, a principal motivação para incluir *rendfam* na equação é que suspeitamos que ela seja correlacionada com *gasto*, e, desse modo, gostaríamos de mantê-la fixa na análise. A hipótese RLM.4 somente exclui a correlação perfeita entre *gasto* e *rendfam* em nossa amostra. Teríamos muito azar se obtivéssemos uma amostra em que os gastos por estudante fossem perfeitamente correlacionados com a renda familiar média. Porém, espera-se que haja alguma correlação — talvez uma quantidade substancial — e certamente ela é permitida.

A maneira mais simples como duas variáveis independentes podem ser perfeitamente correlacionadas é quando uma variável é um múltiplo constante da outra. Isso pode acontecer quando um pesquisador, inadvertidamente, coloca a mesma variável medida em unidades diferentes dentro da equação de regressão. Por exemplo, ao estimar a relação entre consumo e renda, não faz sentido incluir como variáveis independentes a renda mensurada em dólares e a renda mensurada em milhares de dólares. Uma delas é redundante. Que sentido faria manter a renda mensurada em dólares fixa, enquanto a renda mensurada em milhares de dólares varia?

Já sabemos que diferentes funções não-lineares da mesma variável *podem* aparecer entre os regressores. Por exemplo, o modelo $cons = \beta_0 + \beta_1 rend + \beta_2 rend^2 + u$ não viola a hipótese RLM.4: ainda que $x_2 = rend^2$ seja uma função exata de $x_1 = rend$, $rend^2$ não é uma função *linear* de *rend*. Incluir $rend^2$ no modelo é uma maneira útil de generalizar a forma funcional, diferentemente de incluir a renda mensurada em dólares e em milhares de dólares.

O senso comum nos diz para não incluir a mesma variável explicativa medida em diferentes unidades na mesma equação de regressão. Há também maneiras mais sutis de uma variável independente poder ser um múltiplo de outra. Suponha que gostaríamos de estimar uma extensão da função de consumo de elasticidade constante. Poderia parecer natural especificar um modelo tal como

$$\log(cons) = \beta_0 + \beta_1 \log(rend) + \beta_2 \log(rend^2) + u, \qquad \textbf{(3.35)}$$

em que $x_1 = \log(rend)$ e $x_2 = \log(rend^2)$. Usando as propriedades básicas do log natural (veja o Apêndice A, disponível no site de Thomson), $\log(rend^2) = 2 \cdot \log(rend)$. Isto é, $x_2 = 2x_1$, e naturalmente isso é válido para todas as observações na amostra. Isso viola a hipótese RLM.4. Em vez disso, deveríamos incluir $[\log(rend)]^2$, e não $\log(rend^2)$, juntamente com $\log(rend)$. Essa é uma extensão prudente do modelo de elasticidade constante, e veremos como interpretar tais modelos no Capítulo 6.

Outra maneira de as variáveis independentes serem perfeitamente colineares ocorre quando uma variável independente pode ser expressa como uma função linear exata de duas ou mais das outras variáveis independentes. Por exemplo, suponha que queremos estimar o efeito dos gastos de campanha sobre os resultados da campanha. Por simplicidade, assuma que cada eleição tem dois candidatos. Seja *votoA* a percentagem de votos do Candidato A; seja *gastoA* os gastos de campanha do Candidato A; seja *gastoB* os gastos de campanha do Candidato B; e seja *totalgasto* os gastos totais de campanha; todas as últimas três variáveis são medidas em dólares. Pode parecer natural especificar o modelo como

$$votoA = \beta_0 + \beta_1 gastoA + \beta_2 gastoB + \beta_3 totalgasto + u, \qquad \textbf{(3.36)}$$

a fim de isolar os efeitos dos gastos de cada candidato e da quantidade total de gastos. No entanto, esse modelo viola a hipótese RLM.4, porque $x_3 = x_1 + x_2$ por definição. Tentar interpretar essa equação ao estilo *ceteris paribus* revela o problema. Supõe-se que o parâmetro de β_1 na equação (3.36) meça o efeito de um aumento de um dólar nos gastos do Candidato A sobre os votos do Candidato A, mantendo os gastos do Candidato B *e* os gastos totais fixos. Isso é uma tolice, pois, se *gastoB* e *totalgasto* forem mantidos fixos, não podemos aumentar *gastoA*.

A solução para a colinearidade perfeita em (3.36) é simples: retire qualquer uma das três variáveis do modelo. Provavelmente, tiraríamos *totalgasto*, e conseqüentemente o coeficiente de *gastoA* mensuraria o efeito de aumentar os gastos de A sobre a percentagem de votos recebidos por A, mantendo os gastos de B fixos.

O exemplo anterior mostra que a hipótese RLM.4 pode ser violada se não formos cuidadosos ao especificar nosso modelo. Essa hipótese também não se mantém se o tamanho da amostra, n, é muito pequeno em relação ao número de parâmetros que são estimados. No modelo de regressão geral da equação (3.31), há $k + 1$ parâmetros, e RLM.4 não se mantém se $n < k + 1$. Intuitivamente, isso faz sentido: para estimar $k + 1$ parâmetros, necessitamos de pelo menos $k + 1$ observações. Não surpreendentemente, é melhor ter tantas observações quanto possível, algo que veremos em nossos cálculos da variância na Seção 3.4.

No exemplo anterior, se usarmos como variáveis explicativas *gastoA*, *gastoB* e *partA*, em que *partA* = 100·(*gastoA*/*totalgasto*) é a participação percentual dos gastos totais de campanha feitos pelo Candidato A, isso viola a hipótese RLM.4?

Se o modelo for cuidadosamente especificado e $n \geq k + 1$, a hipótese RLM.4 pode não se manter em casos raros devido a um azar ao coletar a amostra. Por exemplo, em uma equação de salários, com educação e experiência como variáveis, poderíamos obter uma amostra aleatória em que cada indivíduo tivesse exatamente duas vezes mais educação que anos de experiência. Esse cenário faria com que a hipótese RLM.4 falhasse, mas isso pode ser considerado muito improvável, a não ser que tenhamos um tamanho de amostra extremamente pequeno.

Agora, estamos prontos para mostrar que, sob essas quatro hipóteses da regressão múltipla, os estimadores de MQO são não-viesados. Como no caso da regressão simples, as esperanças estão condicionadas aos valores das variáveis independentes da amostra, mas não mostraremos esse condicionamento explicitamente.

T E O R E M A 3 . 1 (INEXISTÊNCIA DE VIÉS DE MQO)

Sob as hipóteses RLM.1 a RLM.4,

$$E(\hat{\beta}_j) = \beta_j, j = 0, 1, ..., k, \qquad \textbf{(3.37)}$$

para qualquer valor do parâmetro populacional β_j. Em outras palavras, os estimadores de MQO são estimadores não-viesados dos parâmetros da população.

Em nossos exemplos empíricos anteriores, a hipótese RLM.4 foi satisfeita (visto que fomos capazes de calcular as estimativas de MQO). Além disso, em sua maior parte, as amostras são aleatoriamente escolhidas de uma população bem-definida. Se acreditamos que os modelos especificados estão corretos sob a hipótese fundamental RLM.3, então podemos concluir que MQO é não-viesado nesses exemplos.

Como estamos nos aproximando do ponto em que podemos usar a regressão múltipla no trabalho empírico é útil lembrar o significado de inexistência de viés. É tentador, nos exemplos como o da equação do salário em (3.19), dizer algo como "9,2% é uma estimativa não-viesada do retorno da educação". Como sabemos, uma estimativa não pode ser viesada: uma estimativa é um número fixo, obtido de uma amostra particular, usualmente diferente do parâmetro populacional. Quando dizemos que MQO é não-viesado sob as hipóteses RLM.1 a RLM.4, estamos dizendo que o *procedimento* pelo qual as estimativas de MQO foram obtidas é não-viesado, e tal procedimento é visto como algo aplicado em todas as amostras aleatórias possíveis. Esperamos que tenhamos obtido uma amostra que nos dê uma estimativa próxima do valor da população, mas, infelizmente, isso não pode ser garantido.

Inclusão de Variáveis Irrelevantes em um Modelo de Regressão

Uma questão que podemos dispensar com rapidez razoável é a **inclusão de uma variável irrelevante** ou a **superespecificação do modelo** na análise de regressão múltipla. Isso significa que uma (ou mais) das variáveis independentes está incluída no modelo, embora ela não tenha efeito parcial sobre y na população. (Isto é, seu coeficiente populacional é zero.)

Para ilustrar a questão, suponha que especificamos o modelo como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u, \tag{3.38}$$

e esse modelo satisfaz as hipóteses RLM.1 a RLM.4. Entretanto, x_3 não tem efeito sobre y após x_1 e x_2 terem sido controlados, o que significa que $\beta_3 = 0$. A variável x_3 pode ou não ser correlacionada com x_1 e x_2; o que importa é que, uma vez que x_1 e x_2 estejam controlados, x_3 não tem efeito sobre y. Em termos de esperanças condicionais, $E(y|x_1,x_2,x_3) = E(y|x_1,x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

Como não sabemos que $\beta_3 = 0$, somos inclinados a estimar a equação com x_3:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3. \tag{3.39}$$

Incluímos a variável irrelevante, x_3, em nossa regressão. Qual é o efeito de incluir x_3 em (3.39), quando seu coeficiente no modelo populacional (3.38) é zero? Em termos da inexistência de viés de $\hat{\beta}_1$ e $\hat{\beta}_2$, não há *nenhum efeito*. Essa conclusão não requer nenhuma derivação especial, já que ela decorre imediatamente do Teorema 3.1. Lembre-se, a inexistência de viés significa $E(\hat{\beta}_j) = \beta_j$ para *qualquer* valor de β_j, incluindo $\beta_j = 0$. Assim, podemos concluir que $E(\hat{\beta}_0) = \beta_0$, $E(\hat{\beta}_1) = \beta_1$, $E(\hat{\beta}_2) = \beta_2$ e $E(\hat{\beta}_3) = 0$ (para quaisquer valores de β_0, β_1 e β_2). Mesmo que $\hat{\beta}_3$, por si mesmo, nunca seja exatamente zero, seu valor médio obtido de muitas amostras aleatórias será zero.

A conclusão do exemplo anterior é muito mais geral: incluir uma ou mais variáveis irrelevantes no modelo de regressão múltipla, ou superespecificar o modelo, não afeta a inexistência de viés dos estimadores de MQO. Isso significa que incluir variáveis irrelevantes é inócuo? Não. Como veremos na Seção 3.4, incluir variáveis irrelevantes pode ter efeitos indesejáveis sobre as *variâncias* dos estimadores de MQO.

Viés de Variável Omitida: O Caso Simples

Suponha agora que, em vez de incluir uma variável irrelevante, omitimos uma variável que, realmente, pertence ao modelo verdadeiro (ou populacional). Isso é freqüentemente chamado problema de **excluir uma variável relevante** ou de **subespecificar o modelo**. No Capítulo 2, e anteriormente neste capítulo, afirmamos que esse problema geralmente faz com que os estimadores de MQO sejam viesados. Agora é o momento de mostrar isso explicitamente e, não menos importante, derivar a direção e o tamanho do viés.

Derivar o viés causado ao omitir uma variável importante é um exemplo de **análise de má-especificação**. Iniciaremos com o caso em que o modelo populacional verdadeiro tem duas variáveis explicativas e um termo erro:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \tag{3.40}$$

e assumimos que esse modelo satisfaz as hipóteses RLM.1 a RLM.4.

Suponha que nosso interesse primordial esteja em β_1, o efeito parcial de x_1 sobre y. Por exemplo, y é o salário horário (ou log do salário horário), x_1 é educação e x_2 é uma medida de aptidão inata. A fim de obter um estimador não-viesado de β_1, *deveríamos* computar a regressão de y sobre x_1 e x_2 (o que resulta em estimadores não-viesados de β_0, β_1 e β_2). Entretanto, devido à nossa ignorância ou indisponibilidade de dados, estimamos o modelo *excluindo* x_2. Em outras palavras, executamos somente uma regressão simples de y sobre x_1, obtendo a equação

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1. \tag{3.41}$$

Usamos o símbolo gráfico "~" em vez de "^" para enfatizar que $\tilde{\beta}_1$ vem de um modelo subespecificado.

Ao aprender, pela primeira vez, o problema de variável omitida, pode ser difícil para o estudante distinguir entre o modelo verdadeiro subjacente, (3.40) nesse caso, e o modelo que realmente estimamos, capturado pela regressão em (3.41). Pode parecer bobagem omitir a variável x_2 se ela pertence ao modelo, mas freqüentemente não temos escolha. Por exemplo, suponha que *salárioh* seja determinado pela equação

$$salárioh = \beta_0 + \beta_1 educ + \beta_2 aptid + u. \tag{3.42}$$

Como a aptidão não é observada, estimamos, em vez disso, o modelo

$$salárioh = \beta_0 + \beta_1 educ + v,$$

onde $v = \beta_2 aptid + u$. O estimador de β_1 da regressão simples de *salárioh* sobre *educ* é o que estamos chamando $\tilde{\beta}_1$.

Vamos derivar o valor esperado de $\tilde{\beta}_1$ condicionado aos valores amostrais de x_1 e x_2. Derivar essa esperança não é difícil, pois $\tilde{\beta}_1$ é exatamente o estimador de inclinação de MQO de uma regressão simples, e já estudamos esse estimador extensivamente no Capítulo 2. A diferença aqui é que devemos analisar suas propriedades quando o modelo de regressão simples é mal-especificado devido a uma variável omitida.

Da equação (2.49), podemos expressar $\tilde{\beta}_1$ como

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1) y_i}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2}. \tag{3.43}$$

O próximo passo é o mais importante. Visto que (3.40) é o modelo verdadeiro, escrevemos y para cada observação i como

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i \tag{3.44}$$

(e *não* $y_i = \beta_0 + \beta_1 x_{i1} + u_i$, já que o modelo verdadeiro contém x_2). Seja SQT_1 o denominador em (3.43). Se inserirmos em (3.43) o y_i de (3.44), o numerador em (3.43) passa a ser

$$\begin{aligned} &\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i) \\ &= \beta_1 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 + \beta_2 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i2} + \sum_{i=1}^{n} (x_{i1} - \bar{x}_1) u_i \\ &= \beta_1 \text{SQT}_1 + \beta_2 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i2} + \sum_{i=1}^{n} (x_{i1} - \bar{x}_1) u_i. \end{aligned} \tag{3.45}$$

Se dividirmos (3.45) por SQT_1, considerarmos a esperança condicionada aos valores das variáveis independentes e usarmos $\text{E}(u_i) = 0$, obteremos

$$\text{E}(\tilde{\beta}_1) = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i2}}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2}. \tag{3.46}$$

Assim, $\text{E}(\tilde{\beta}_1)$ não é, geralmente, igual a β_1: $\tilde{\beta}_1$ é viesado para β_1.

A razão que multiplica β_2 em (3.46) tem uma interpretação simples: ela é exatamente o coeficiente de inclinação da regressão de x_2 sobre x_1, utilizando nossa amostra de variáveis independentes, que pode ser escrita como

$$\tilde{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1. \tag{3.47}$$

Como estamos condicionados aos valores amostrais de ambas as variáveis independentes, $\tilde{\delta}_1$ não é aleatório aqui. Portanto, podemos escrever (3.46) como

$$\text{E}(\tilde{\beta}_1) = \beta_1 + \beta_2 \tilde{\delta}_1, \tag{3.48}$$

o que implica que o viés em $\tilde{\beta}_1$ é $E(\tilde{\beta}_1) - \beta_1 = \beta_2\tilde{\delta}_1$. Essa expressão é chamada freqüentemente de **viés de variável omitida**.

Da equação (3.48), vemos que há dois casos em que $\tilde{\beta}_1$ é não-viesado. O primeiro é bastante óbvio: se $\beta_2 = 0$ — de modo que x_2 não aparece no modelo verdadeiro (3.40) —, então $\tilde{\beta}_1$ é não-viesado. Já sabemos isso da análise de regressão simples do Capítulo 2. O segundo caso é mais interessante. Se $\tilde{\delta}_1 = 0$, então $\tilde{\beta}_1$ é não-viesado para β_1, mesmo se $\beta_2 \neq 0$.

Como $\tilde{\delta}_1$ é a covariância amostral entre x_1 e x_2 sobre a variância amostral de x_1, $\tilde{\delta}_1 = 0$ se, e somente se, x_1 e x_2 forem não-correlacionados na amostra. Assim, temos a importante conclusão de que, se x_1 e x_2 forem não-correlacionados na amostra, então $\tilde{\beta}_1$ é não-viesado. Isso não é surpreendente: na Seção 3.2, mostramos que o estimador da regressão simples $\tilde{\beta}_1$ e o estimador da regressão múltipla $\hat{\beta}_1$ são iguais quando x_1 e x_2 forem não-correlacionados na amostra. [Podemos também mostrar que $\tilde{\beta}_1$ é não-viesado sem condicionar a x_{i2} se $E(x_2|x_1) = E(x_2)$; então, para a estimação de β_1, deixar x_2 no termo erro não viola a hipótese de média condicional zero do erro, uma vez que ajustamos o intercepto.]

Quando x_1 e x_2 forem correlacionados, $\tilde{\delta}_1$ tem o mesmo sinal da correlação entre x_1 e x_2: $\tilde{\delta}_1 > 0$ se x_1 e x_2 forem positivamente correlacionados, e $\tilde{\delta}_1 < 0$ se x_1 e x_2 forem negativamente correlacionados. O sinal do viés em $\tilde{\beta}_1$ depende tanto do sinal de β_2 como de $\tilde{\delta}_1$ e está resumido na Tabela 3.2 para os quatro casos possíveis quando há viés. A Tabela 3.2 justifica um estudo cuidadoso. Por exemplo, o viés em $\tilde{\beta}_1$ é positivo se $\beta_2 > 0$ (x_2 tem um efeito positivo sobre y) e x_1 e x_2 são positivamente correlacionados; o viés é negativo se $\beta_2 > 0$ e x_1 e x_2 são negativamente correlacionados, e assim por diante.

Tabela 3.2

Sumário do Viés em $\tilde{\beta}_1$ quando x_2 é Omitida na Estimação da Equação (3.40)

	$\mathbf{Corr}(x_1,x_2) > 0$	$\mathbf{Corr}(x_1,x_2) < 0$
$\beta_2 > 0$	viés positivo	viés negativo
$\beta_2 < 0$	viés negativo	viés positivo

A Tabela 3.2 resume a direção do viés, mas o tamanho do viés também é muito importante. Um viés pequeno de qualquer dos dois sinais não precisa ser causa de preocupação. Por exemplo, se o retorno da educação formal é de 8,6% na população e o viés do estimador de MQO é de 0,1% (um décimo de um ponto percentual), então não precisaríamos ficar muito preocupados. De outro lado, um viés da ordem de três pontos percentuais seria muito mais sério. O tamanho do viés é determinado pelos tamanhos de β_2 e $\tilde{\delta}_1$.

Na prática, como β_2 é um parâmetro populacional desconhecido, não podemos estar certos se β_2 é positivo ou negativo. Entretanto, temos geralmente uma boa idéia sobre a direção do efeito parcial de x_2 sobre y. Além disso, ainda que o sinal da correlação entre x_1 e x_2 não possa ser conhecido se x_2 não é observado, em muitos casos, podemos fazer uma suposição criteriosa sobre se x_1 e x_2 são positiva ou negativamente correlacionados.

Na equação do salário (3.42), por definição, mais aptidão conduz a uma produtividade maior e, portanto, a salários maiores: $\beta_2 > 0$. Há também razões para acreditar que *educ* e *aptid* sejam positivamente correlacionados: em média, indivíduos com mais aptidão inata escolhem níveis maiores de educação formal. Assim, as estimativas de MQO da equação de regressão simples do $salárioh = \beta_0 + \beta_1 educ + v$ são, *em média*, muito grandes. Isso não significa que a estimativa obtida de nossa amostra seja enorme. Somente podemos dizer que, se coletarmos muitas amostras aleatoriamente e obtivermos as estimativas da regressão simples a cada vez, a média dessas estimativas será maior que β_1.

EXEMPLO 3.6

(Equação do Salário Horário)

Suponha que o modelo $\log(salárioh) = \beta_0 + \beta_1 educ + \beta_2 aptid + u$ satisfaça as hipóteses RLM.1 a RLM.4. O conjunto de dados no arquivo WAGE1.RAW não contém dados sobre aptidão, de modo que estimamos β_1 a partir da regressão simples

$$\widetilde{\log(salárioh)} = 0{,}584 + 0{,}083\, educ$$
$$n = 526,\ R^2 = 0{,}186.$$

Esse é somente o resultado de uma única amostra, de modo que não podemos dizer que 0,083 é maior que β_1; o retorno verdadeiro da educação poderia ser menor ou maior que 8,3% (nunca saberemos com certeza). Entretanto, sabemos que a média dos estimadores de todas as amostras aleatórias seria bastante grande.

Como um segundo exemplo, suponha que, no nível fundamental do ensino, a nota média dos estudantes de um exame padronizado seja determinado por

$$notmed = \beta_0 + \beta_1 gasto + \beta_2 taxpob + u,$$

em que *gasto* é o gasto público por estudante, e *taxpob* é a taxa de pobreza das crianças da escola. Usando dados do distrito da escola, temos somente observações da percentagem de estudantes com uma nota de aprovação e gastos públicos por estudante; não temos informações sobre taxas de pobreza. Assim, estimamos β_1 a partir da regressão simples de *notmed* sobre *gasto*.

Podemos obter, uma vez mais, o viés provável em $\tilde{\beta}_1$. Primeiro, β_2 é provavelmente negativo: há ampla evidência de que crianças que vivem na pobreza têm, em média, notas mais baixas em testes padronizados. Segundo, o gasto público médio por estudante é, provavelmente, negativamente correlacionado com a taxa de pobreza: quanto maior a taxa de pobreza menor o gasto público médio por estudante, de modo que $\text{Corr}(x_1, x_2) < 0$. De acordo com a Tabela 3.2, $\tilde{\beta}_1$ terá um viés positivo. Essa observação tem implicações importantes. Pode ser que o efeito verdadeiro do gasto público fosse zero; isto é, $\beta_1 = 0$. Entretanto, a estimativa de β_1 da regressão simples será, geralmente, maior que zero, e isso poderia nos levar a concluir que os gastos públicos são importantes quando eles não são.

Ao ler e ao fazer trabalhos empíricos em economia, é importante dominar a terminologia associada aos estimadores viesados. No contexto de omissão de uma variável do modelo (3.40), se $\text{E}(\tilde{\beta}_1) > \beta_1$, então dizemos que $\tilde{\beta}_1$ tem um **viés para cima**. Quando $\text{E}(\tilde{\beta}_1) < \beta_1$, $\tilde{\beta}_1$ tem um **viés para baixo**. Essas definições são as mesmas, seja β_1 positivo ou negativo. A expressão **viesado para zero** refere-se aos casos em que $\text{E}(\tilde{\beta}_1)$ está mais próxima de zero do que de β_1. Portanto, se β_1 for positivo, $\tilde{\beta}_1$ será viesado para zero se ele tiver um viés para baixo. De outro lado, se $\beta_1 < 0$, $\tilde{\beta}_1$ será viesado para zero se ele tiver um viés para cima.

Viés de Variável Omitida: Casos mais Gerais

É mais difícil derivar o sinal do viés de variável omitida quando há múltiplos regressores no modelo estimado. Devemos lembrar que a correlação entre uma única variável explicativa e o erro resulta, geralmente, em *todos* os estimadores de MQO serem viesados. Por exemplo, suponha que o modelo populacional

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \qquad \textbf{(3.49)}$$

satisfaça as hipóteses RLM.1 a RLM.4. No entanto, omitimos x_3 e estimamos o modelo como

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2. \qquad \textbf{(3.50)}$$

Agora suponha que x_2 e x_3 sejam não-correlacionados, mas que x_1 é correlacionado com x_3. Em outras palavras, x_1 é correlacionado com a variável omitida, mas x_2 não é. É tentador pensar que, embora provavelmente $\tilde{\beta}_1$ seja viesado com base na derivação da subseção anterior, $\tilde{\beta}_2$ seja não-viesado, pois x_2 é não-correlacionado com x_3. Infelizmente, esse *não* é, geralmente, o caso: normalmente, tanto $\tilde{\beta}_1$ como $\tilde{\beta}_2$ serão viesados. A única exceção a isso ocorre quando x_1 e x_2 também são não-correlacionados.

Mesmo em um modelo razoavelmente simples como o apresentado, pode ser difícil obter a direção do viés em $\tilde{\beta}_1$ e $\tilde{\beta}_2$. Isso se deve ao fato de que x_1, x_2 e x_3 podem estar correlacionados aos pares. Entretanto, uma aproximação é, freqüentemente, útil na prática. Se assumirmos que x_1 e x_2 são não-correlacionados, podemos estudar o viés em $\tilde{\beta}_1$ como se x_2 estivesse ausente dos modelos populacional e estimado. De fato, quando x_1 e x_2 são não-correlacionados, pode-se mostrar que

$$\mathrm{E}(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1) x_{i3}}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2}$$

Essa equação é exatamente igual a (3.46), mas β_3 substitui β_2, e x_3 substitui x_2. Portanto, o viés em $\tilde{\beta}_1$ é obtido ao se substituir β_2 por β_3 e x_2 por x_3 na Tabela 3.2. Se $\beta_3 > 0$ e $\mathrm{Corr}(x_1, x_3) > 0$, o viés em $\tilde{\beta}_1$ é positivo, e assim por diante.

Como um exemplo, suponha que acrescentamos *exper* ao modelo do salário:

$$salárioh = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 aptid + u.$$

Se *aptid* for omitida do modelo, ambos os estimadores de β_1 e β_2 serão viesados, mesmo se assumirmos que *exper* é não-correlacionado com *aptid*. Estamos principalmente interessados no retorno da educação formal, de modo que seria bom se pudéssemos concluir que $\tilde{\beta}_1$ tem um viés para cima ou para baixo devido à omissão da aptidão. Essa conclusão não é possível sem hipóteses adicionais. Como uma *aproximação*, suponhamos que, além de *exper* e *aptid* serem não-correlacionadas, *educ* e *exper* também sejam não-correlacionadas. (Na realidade, elas são negativamente correlacionadas.) Como $\beta_3 > 0$ e *educ* e *aptid* são positivamente correlacionadas, $\tilde{\beta}_1$ teria um viés para cima, exatamente como se *exper* não estivesse no modelo.

O raciocínio usado no exemplo anterior é, muitas vezes, compreendido como um guia aproximado para obter o viés provável dos estimadores em modelos mais complicados. Geralmente, o foco está

na relação entre uma variável explicativa particular, por exemplo x_1, e o fator omitido fundamental. Estritamente falando, ignorar todas as outras variáveis explicativas é uma prática válida somente quando cada uma delas é não-correlacionada com x_1, mas essa ainda é uma orientação útil. O Apêndice 3A (disponível no site de Thomson) contém uma análise mais cuidadosa do viés de variável omitida com múltiplas variáveis explicativas.

3.4 A VARIÂNCIA DOS ESTIMADORES DE MQO

Obteremos, agora, a variância dos estimadores de MQO, de modo que, além de conhecermos as tendências centrais dos $\hat{\beta}_j$, também teremos uma medida da dispersão de sua distribuição amostral. Antes de encontrarmos as variâncias, vamos adicionar uma hipótese de homoscedasticidade, como no Capítulo 2. Fazemos isso por duas razões. Primeira, ao impor a hipótese de variância constante do erro, as fórmulas são simplificadas. Segunda, veremos na Seção 3.5 que MQO tem uma propriedade importante de eficiência se acrescentamos a hipótese de homoscedasticidade.

No arcabouço da regressão múltipla, a homoscedasticidade é expressa como a seguir:

HIPÓTESE RLM.5 (HOMOSCEDASTICIDADE)
$\text{Var}(u|x_1,...,x_k) = \sigma^2$.

A hipótese RLM.5 significa que a variância do termo erro, u, condicionada às variáveis explicativas, é a *mesma* para todas as combinações de resultados das variáveis explicativas. Se essa hipótese é violada, o modelo exibe heteroscedasticidade, exatamente como no caso de duas variáveis.

Na equação

$$salárioh = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 perm + u,$$

a homoscedasticidade requer que a variância do erro não-observado u não dependa dos níveis de educação, experiência ou permanência. Isto é,

$$\text{Var}(u|educ,exper,perm) = \sigma^2.$$

Se a variância varia com qualquer uma das três variáveis explicativas, então a heteroscedasticidade está presente.

As hipóteses RLM.1 a RLM.5 são, em conjunto, conhecidas como as **hipóteses de Gauss-Markov** (para a regressão de corte transversal). Até agora, nossas asserções sobre as hipóteses são adequadas somente quando aplicadas à análise de corte transversal com amostragem aleatória. Como veremos, as hipóteses de Gauss-Markov para a análise de séries de tempo — e para outras situações, como a análise de dados de painel — são mais difíceis de se manterem, embora haja muitas similaridades.

Na discussão a seguir, usaremos o símbolo $\boldsymbol{x}$ para representar o conjunto de todas as variáveis independentes, $(x_1, ..., x_k)$. Assim, na regressão do salário horário com *educ*, *exper* e *perm* como variáveis independentes, $\boldsymbol{x}$ = (*educ*, *exper*, *perm*). Conseqüentemente, podemos escrever as hipóteses RLM.1 e RLM.3 como

$$\text{E}(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_3 x_3,$$

e a hipótese RLM.5 é a mesma que $\text{Var}(y|\boldsymbol{x}) = \sigma^2$. Expressar as hipóteses desse modo ilustra como a hipótese RLM.5 difere muito da hipótese RLM.3. Esta diz que o valor esperado de y, dado $\boldsymbol{x}$, é linear nos parâmetros, mas ele certamente depende de $x_1, x_2, ..., x_k$. A hipótese RLM.5 diz que a variância de y, dado $\boldsymbol{x}$, *não* depende dos valores das variáveis independentes.

Podemos obter, agora, as variâncias dos $\hat{\beta}_j$, que uma vez mais, estão condicionadas aos valores amostrais das variáveis independentes. A prova está no apêndice deste capítulo.

TEOREMA 3.2 (VARIÂNCIAS AMOSTRAIS DOS ESTIMADORES DE INCLINAÇÃO DE MQO)

Sob as hipóteses RLM.1 a RLM.5, condicionadas aos valores amostrais das variáveis independentes,

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\text{SQT}_j(1 - R_j^2)}, \tag{3.51}$$

para $j = 1, 2, ..., k$, em que $\text{SQT}_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$ é a variação amostral total em x_j, e R_j^2 é o R-quadrado da regressão x_j sobre todas as outras variáveis independentes (incluindo um intercepto).

Antes de estudarmos a equação (3.51) mais detalhadamente, é importante saber que todas as hipóteses de Gauss-Markov são usadas na obtenção dessa fórmula. Embora não necessitemos da hipótese de homoscedasticidade para concluir que MQO é não-viesado, precisamos dela para validar a equação (3.51).

O tamanho de $\text{Var}(\hat{\beta}_j)$ é importante na prática. Uma variância maior significa um estimador menos preciso, e isso se traduz em intervalos de confiança maiores e testes de hipóteses menos acurados (como veremos no Capítulo 4). Na próxima subseção, discutiremos os elementos que compreendem (3.51).

Os Componentes das Variâncias de MQO: Multicolinearidade

A equação (3.51) mostra que a variância de $\hat{\beta}_j$ depende de três fatores: σ^2, SQT_j e R_j^2. Lembre-se de que o índice j representa simplesmente qualquer uma das variáveis independentes (como a educação ou a taxa de pobreza). Agora, vamos considerar cada um dos fatores que afetam $\text{Var}(\hat{\beta}_j)$.

A VARIÂNCIA DO ERRO, σ^2. Da equação (3.51), um σ^2 maior significa variâncias maiores dos estimadores de MQO. Isso não é totalmente surpreendente: mais "ruído" na equação (um σ^2 maior) torna mais difícil estimar o efeito parcial de qualquer uma das variáveis independentes sobre y, e isso é refletido nas variâncias maiores dos estimadores de inclinação de MQO. Visto que σ^2 é uma característica da população, ele não tem nada a ver com o tamanho da amostra. Ele é o componente de (3.51) que é desconhecido. Veremos mais adiante como obter um estimador não-viesado de σ^2.

Para uma dada variável dependente y, há de fato somente uma maneira de reduzir a variância do erro, que é adicionar mais variáveis explicativas à equação (retirar alguns fatores do termo erro). Isso nem sempre é possível, nem sempre é desejável, por razões discutidas posteriormente neste capítulo.

A VARIAÇÃO AMOSTRAL TOTAL EM x_j, SQT_j. Da equação (3.51), vemos que quanto maior a variação total em x_j, menor é $\text{Var}(\hat{\beta}_j)$, Assim, tudo o mais sendo igual para estimar $\hat{\beta}_j$ preferimos ter tanta variação amostral em x_j quanto possível. Já descobrimos isso no caso da regressão simples, no Capítulo 2. Embora raramente seja possível escolher os valores amostrais das variáveis independentes, *há* uma maneira de aumentar a variação amostral em cada uma das variáveis independentes: aumentar o tamanho da amostra. De fato, na amostragem aleatória de uma população, SQT_j aumenta sem limite quando o tamanho da amostra torna-se maior. Esse é o componente da variância que depende sistematicamente do tamanho da amostra.

Quando SQT_j é pequeno, $\text{Var}(\hat{\beta}_j)$ pode ficar muito grande, mas um SQT_j pequeno não é uma violação da hipótese RLM.4. Tecnicamente, quando SQT_j tende a zero, $\text{Var}(\hat{\beta}_j)$ aproxima-se do infinito. O caso extremo de nenhuma variação amostral em x_j, $SQT_j = 0$, não é permitido pela hipótese RLM.4.

AS RELAÇÕES LINEARES ENTRE AS VARIÁVEIS INDEPENDENTES, R_j^2. O termo R_j^2 na equação (3.51) é o mais difícil dos três componentes de se entender. Esse termo não aparece na análise de regressão simples porque há somente uma variável independente em tal caso. É importante compreender que esse R-quadrado é distinto do R-quadrado da regressão de y sobre $x_1, x_2, \ldots, x_k$: R_j^2 é obtido de uma regressão que envolve somente as variáveis independentes do modelo original, em que x_j desempenha o papel de uma variável dependente.

Considere, primeiro, o caso $k = 2$: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$. Então, $\text{Var}(\hat{\beta}_1) = \sigma^2/[SQT_1(1 - R_1^2)]$, em que R_1^2 é o R-quadrado da regressão simples de x_1 sobre x_2 (e um intercepto, como sempre). Como o R-quadrado mede o grau de ajuste, um valor de R_1^2 próximo de um indica que x_2 explica bastante da variação de x_1 na amostra. Isso significa que x_1 e x_2 são altamente correlacionados.

Quando R_1^2 cresce em direção a um, $\text{Var}(\hat{\beta}_1)$ torna-se maior. Assim, um grau elevado de relação linear entre x_1 e x_2 pode levar a variâncias grandes dos estimadores de inclinação de MQO. (Um argumento similar se aplica a $\hat{\beta}_2$.) Veja a Figura 3.1 para a relação entre $\text{Var}(\hat{\beta}_1)$ e o R-quadrado da regressão de x_1 sobre x_2.

No caso geral, R_j^2 é a proporção da variação total de x_j que pode ser explicada pelas *outras* variáveis independentes que aparecem na equação. Para dados σ^2 e SQT_j, a menor $\text{Var}(\hat{\beta}_j)$ é obtida quando $R_j^2 = 0$, que ocorre se, e somente se, x_j tem correlação amostral zero com *cada uma das outras* variáveis independentes. Esse é o melhor caso para estimar β_j, mas é raramente encontrado.

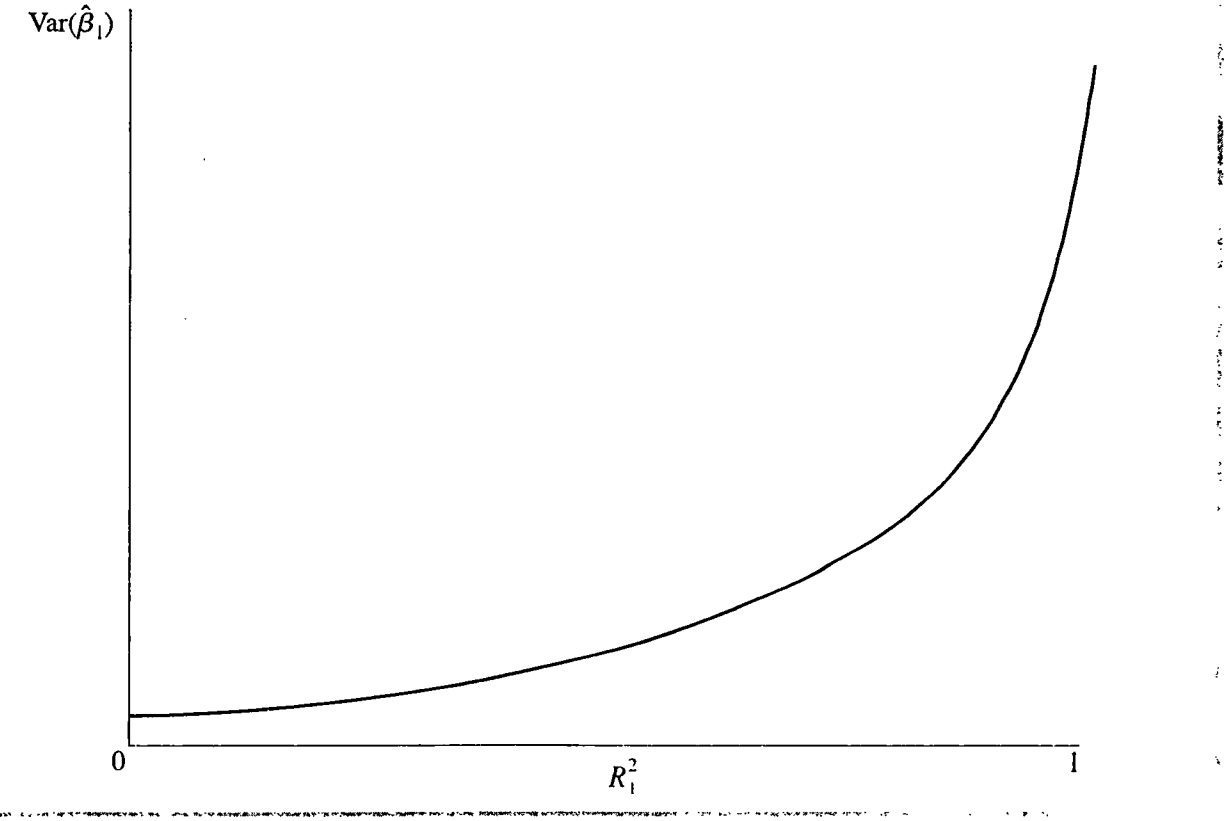
O outro caso extremo, $R_j^2 = 1$, é excluído pela hipótese RLM.4, pois $R_j^2 = 1$ significa que, na amostra, x_j é uma combinação linear *perfeita* de algumas das outras variáveis independentes da regressão. Um caso mais relevante é quando R_j^2 está "próximo" de um. Da equação (3.51) e da Figura 3.1, vemos que isso pode fazer com que $\text{Var}(\hat{\beta}_j)$ seja grande: $\text{Var}(\hat{\beta}_j) \to \infty$ quando $R_j^2 \to 1$. Correlação alta (mas não perfeita) entre duas ou mais variáveis independentes é chamada **multicolinearidade**.

Antes de discutirmos mais a questão da multicolinearidade, é importante que uma coisa esteja bem clara: um caso em que R_j^2 está próximo de um *não* é uma violação da hipótese RLM.4.

Como a multicolinearidade não viola nenhuma de nossas hipóteses, o "problema" da multicolinearidade não é, de fato, bem definido. Ao dizer que a multicolinearidade surge ao estimarmos β_j, quando R_j^2 está "próximo" de um, colocamos "próximo" dentro de aspas porque não há um número absoluto que podemos citar para concluir que a multicolinearidade é um problema. Por exemplo, $R_j^2 = 0{,}9$ significa que 90% da variação amostral em x_j pode ser explicada pelas outras variáveis independentes do modelo de regressão. Inquestionavelmente, isso significa que x_j tem uma forte relação linear com as outras variáveis independentes. No entanto, se isso se traduz em uma $\text{Var}(\hat{\beta}_j)$ que é grande demais para ser útil, depende dos tamanhos de σ^2 e SQT_j. Como veremos no Capítulo

4, para a inferência estatística, o que essencialmente importa é quanto $\hat{\beta}_j$ é grande com relação a seu desvio-padrão.

Figura 3.1
$\text{Var}(\hat{\beta}_j)$ como uma função de R_j^2.



Assim como um valor grande de R_j^2 pode causar uma $\text{Var}(\hat{\beta}_j)$ grande, um valor pequeno de SQT_j também pode fazer com que $\text{Var}(\hat{\beta}_j)$ seja grande. Portanto, um tamanho pequeno da amostra pode também levar a variâncias amostrais grandes. Preocupar-se com graus elevados de correlação entre variáveis independentes da amostra não é, de fato, diferente de se preocupar com um tamanho pequeno da amostra: ambos funcionam para aumentar $\text{Var}(\hat{\beta}_j)$.

O famoso econometrista da Universidade de Wisconsin Arthur Goldberger, reagindo à obsessão dos econometristas pela multicolinearidade, criou (jocosamente) o termo **micronumerosidade**, que ele define como o "problema do tamanho pequeno da amostra". [Para uma discussão interessante sobre multicolinearidade e micronumerosidade, veja Goldberger (1991).]

Embora o problema da multicolinearidade não possa ser claramente definido, uma coisa é clara: tudo mais sendo igual, para estimar β_j, é melhor ter menos correlação entre x_j e as outras variáveis independentes. Essa observação muitas vezes leva a uma discussão de como "resolver" o problema da multicolinearidade. Nas ciências sociais, em que somos geralmente coletores passivos de dados, não há uma boa maneira de reduzir as variâncias dos estimadores não-viesados que não seja coletar mais dados. Para um determinado conjunto de dados, podemos tentar, num esforço para reduzir a multicolinearidade, suprimir outras variáveis independentes do modelo. Infelizmente, suprimir uma variável que pertence ao modelo populacional pode levar viés, como vimos na Seção 3.3.

Neste ponto, talvez um exemplo ajude a esclarecer algumas das questões aqui levantadas relativas à multicolinearidade. Suponha que estamos interessados em estimar o efeito de várias categorias de

despesas de escolas sobre o desempenho de estudantes. É provável que as despesas com salários de professores, materiais institucionais, atletismo etc. estejam altamente correlacionadas: escolas mais ricas tendem a gastar mais com tudo, e escolas mais pobres gastam menos com tudo. Não surpreendentemente, pode ser difícil estimar o efeito de qualquer categoria de despesa específica sobre o desempenho dos estudantes quando há pouca variação em uma categoria que não pode ser, em grande medida, explicada por variações das outras categorias de despesas (isso leva a um R_j^2 alto para cada uma das variáveis de despesas). Esses problemas de multicolinearidade podem ser mitigados ao coletar mais dados mas assim, em certo sentido, nós mesmos nos impusemos o problema: estamos formulando questões que podem ser sutis demais para que os dados disponíveis as respondam com alguma precisão. Provavelmente, podemos fazer algo muito melhor mudando o escopo da análise e agrupando todas as categorias de despesa em uma única categoria, desde que não mais estivéssemos tentando estimar o efeito parcial de cada categoria separadamente.

Outro ponto importante é que um elevado grau de correlação entre certas variáveis independentes pode ser irrelevante no que diz respeito a quão bem podemos estimar outros parâmetros do modelo. Por exemplo, considere um modelo com três variáveis independentes:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

em que x_2 e x_3 são altamente correlacionados. Então, $\text{Var}(\hat{\beta}_2)$ e $\text{Var}(\hat{\beta}_3)$ podem ser grandes. Mas o valor da correlação entre x_2 e x_3 não tem efeito direto sobre $\text{Var}(\hat{\beta}_1)$. De fato, se x_1 é não-correlacionado com x_2 e x_3, então $R_1^2 = 0$ e $\text{Var}(\hat{\beta}_1) = \sigma^2/\text{SQT}_1$, independentemente da quantia de correlação existir entre x_2 e x_3. Se β_1 é o parâmetro de interesse, realmente não devemos nos preocupar com o valor da correlação entre x_2 e x_3.

Suponha que você postula um modelo que explica a nota do exame final em termos da freqüência às aulas. Assim, a variável dependente é a nota do exame final, e a principal variável explicativa é o número de aulas freqüentadas. A fim de controlar as aptidões dos estudantes e pelos esforços fora da sala de aula, você inclui entre as variáveis explicativas a nota acumulada durante todo o curso, a nota do teste de avaliação de conhecimentos para ingresso em curso superior e as medidas do desempenho do estudante no ensino médio. Alguém diz: "Você não pode esperar aprender nada com esse exercício, pois todas essas variáveis são, provavelmente, altamente colineares". Qual seria sua resposta?

A observação anterior é importante porque os economistas freqüentemente incluem muitas variáveis de controle a fim de isolar o efeito causal de uma variável particular. Por exemplo, ao olhar para a relação entre as taxas de aprovação de empréstimos e a percentagem de minorias em uma região, poderíamos incluir variáveis como renda média, valor médio das moradias, medidas de inadimplência, e assim por diante, pois esses fatores precisam ser considerados a fim de se extrair conclusões causais sobre a discriminação. Renda, preços de moradia e inadimplência são, geralmente, altamente correlacionados entre si. No entanto, correlações altas entre essas variáveis não tornam mais difícil determinar os efeitos da discriminação.

Variâncias em Modelos Mal Especificados

A escolha de incluir ou não uma variável particular em um modelo de regressão pode ser feita ao analisar o dilema entre viés e variância. Na Seção 3.3 derivamos o viés produzido pela omissão de uma variável relevante quando o modelo verdadeiro contém duas variáveis explicativas. Vamos continuar a análise desse modelo comparando as variâncias dos estimadores de MQO.

Escreva o modelo populacional verdadeiro que satisfaz as hipóteses de Gauss-Markov, como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

Consideremos dois estimadores de β_1. O estimador $\hat{\beta}_1$ é proveniente da regressão múltipla

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2. \tag{3.52}$$

Em outras palavras, incluímos x_2, juntamente com x_1, no modelo de regressão. O estimador $\tilde{\beta}_1$ é obtido ao omitir x_2 do modelo e computando uma regressão simples de y sobre x_1:

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1. \tag{3.53}$$

Quando $\beta_2 \neq 0$, a equação (3.53) exclui uma variável relevante do modelo e, como vimos na Seção 3.3, isso induz um viés em $\tilde{\beta}_1$, a não ser que x_1 e x_2 sejam não-correlacionados. De outro lado, $\hat{\beta}_1$ é não-viesado para β_1, para qualquer valor de β_2, incluindo $\beta_2 = 0$. Segue-se que, se o viés for usado como único critério, $\hat{\beta}_1$ é preferível a $\tilde{\beta}_1$.

A conclusão de que $\hat{\beta}_1$ é sempre preferível a $\tilde{\beta}_1$ não se sustenta quando trazemos a variância para dentro da análise. Condicionando aos valores de x_1 e x_2 na amostra, temos, de (3.51),

$$\text{Var}(\hat{\beta}_1) = \sigma^2/[\text{SQT}_1(1 - R_1^2)], \tag{3.54}$$

em que SQT_1 é a variação total em x_1, e R_1^2 é o R-quadrado da regressão de x_1 sobre x_2. Além disso, uma simples modificação da prova para a regressão de duas variáveis do Capítulo 2 mostra que

$$\text{Var}(\tilde{\beta}_1) = \sigma^2/\text{SQT}_1. \tag{3.55}$$

Comparar (3.55) a (3.54) mostra que $\text{Var}(\tilde{\beta}_1)$ é sempre *menor* que $\text{Var}(\hat{\beta}_1)$, a menos que x_1 e x_2 sejam não-correlacionados na amostra, caso em que os dois estimadores $\tilde{\beta}_1$ e $\hat{\beta}_1$ são os mesmos. Ao assumir que x_1 e x_2 são não-correlacionados, podemos ter as seguintes conclusões:

1. Quando $\beta_2 \neq 0$, $\tilde{\beta}_1$ é viesado, $\hat{\beta}_1$ é não-viesado e $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$.
2. Quando $\beta_2 = 0$, $\tilde{\beta}_1$ e $\hat{\beta}_1$ são ambos não-viesados e $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$.

Da segunda conclusão, é claro que $\tilde{\beta}_1$ é preferido se $\beta_2 = 0$. Intuitivamente, se x_2 não tem um efeito parcial sobre y, incluí-lo no modelo pode somente exacerbar o problema da multicolinearidade, o que

leva a um estimador menos eficiente de β_1. O custo de incluir uma variável irrelevante no modelo é uma variância maior do estimador de β_1.

O caso em que $\beta_2 \neq 0$ é mais difícil. Omitindo x_2 do modelo leva a um estimador viesado de β_1. Tradicionalmente, econometristas têm sugerido comparar o tamanho provável do viés devido à omissão de x_2 com a redução na variância — resumida no tamanho de R_1^2 — para decidir se x_2 deve ser incluído. Entretanto, quando $\beta_2 \neq 0$, há duas razões favoráveis para incluir x_2 no modelo. A mais importante delas é que qualquer viés em $\tilde{\beta}_1$ não diminui quando o tamanho da amostra cresce; de fato, o viés não segue, necessariamente, qualquer padrão. Portanto, podemos em geral pensar o viés como mais ou menos o mesmo para qualquer tamanho de amostra. De outro lado, $\text{Var}(\tilde{\beta}_1)$ e $\text{Var}(\hat{\beta}_1)$ tendem a zero quando n torna-se grande, o que significa que a multicolinearidade induzida pela adição de x_2 torna-se menos importante quando o tamanho da amostra cresce. Em amostras grandes, preferiríamos $\hat{\beta}_1$.

A outra razão para preferir $\hat{\beta}_1$ é mais sutil. A fórmula da variância em (3.55) está condicionada aos valores de x_1 e x_2 na amostra, o que oferece o melhor cenário para $\tilde{\beta}_1$. Quando $\beta_2 \neq 0$, a variância de $\tilde{\beta}_1$ condicionada somente a x_1 é maior que aquela apresentada em (3.55). Intuitivamente, quando $\beta_2 \neq 0$ e x_2 é excluído do modelo, a variância do erro aumenta porque o erro efetivamente contém parte de x_2. Mas, (3.55) ignora o aumento da variância do erro porque ela trata ambos os regressores como não-aleatórios. Uma discussão completa de quais variáveis independentes deveriam ser condicionadas nos desviaria demais de nosso caminho. É suficiente dizer que (3.55) é bastante generosa quando ela aparece para medir a precisão de $\tilde{\beta}_1$.

Estimação de σ^2: Os Erros-Padrão dos Estimadores de MQO

Vamos, agora, mostrar como escolher um estimador não-viesado de σ^2, que nos permitirá obter estimadores não-viesados de $\text{Var}(\tilde{\beta}_j)$.

Como $\sigma^2 = \text{E}(u^2)$, um "estimador" não-viesado de σ^2 é a média amostral dos erros quadrados: $n^{-1}\sum_{i=1}^{n} u_i^2$. Infelizmente, esse não é um estimador verdadeiro, pois não observamos os u_i. Não obstante, lembre-se de que os erros podem ser escritos como $u_i = y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_k x_{ik}$, e a razão real de não observarmos os u_i é que não conhecemos os β_j. Quando substituímos cada β_j por seu estimador de MQO, obtemos os resíduos de MQO:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}.$$

Parece natural estimar σ^2 ao substituir u_i por $\hat{u}_i$. No caso da regressão simples, vimos que isso leva a um estimador viesado. O estimador não-viesado de σ^2 no caso geral da regressão múltipla é

$$\hat{\sigma}^2 = \left(\sum_{i=1}^{n} u_i^2\right) \Big/ (n - k - 1) = \text{SQR}/(n - k - 1). \qquad \textbf{(3.56)}$$

Já encontramos esse estimador no caso $k = 1$ da regressão simples.

O termo $n - k - 1$ em (3.56) representa os **graus de liberdade** (*gl*) do problema geral de MQO com n observações e k variáveis independentes. Como há $k + 1$ parâmetros em um modelo de regressão com k variáveis independentes e um intercepto, podemos escrever

$$gl = n - (k + 1)$$
$$= \text{(número de observações)} - \text{(número de parâmetros estimados)} \quad (3.57)$$

Essa é a maneira mais fácil de calcular os graus de liberdade em uma aplicação particular: conte o número de parâmetros, incluindo o intercepto, e subtraia esse valor do número de observações. (No caso raro em que o intercepto não é estimado, o número de parâmetros diminui em um.)

Tecnicamente, a divisão por $n - k - 1$ em (3.56) é proveniente do fato de que o valor esperado da soma dos resíduos quadrados é $E(SQR) = (n - k - 1)\sigma^2$. Intuitivamente, podemos entender porque o ajustamento de graus de liberdade é necessário ao retornarmos às condições de primeira ordem dos estimadores de MQO. Elas podem ser escritas como $\sum_{i=1}^{n} u_i = 0$ e $\sum_{i=1}^{n} x_{ij}\hat{u}_i = 0$, onde $j = 1, 2, ..., k$. Assim, na obtenção dos estimadores de MQO, $k + 1$ restrições são impostas sobre os resíduos de MQO. Isso significa que, dados $n - (k + 1)$ dos resíduos, os $k + 1$ resíduos restantes são conhecidos: há somente $n - (k + 1)$ graus de liberdade nos resíduos. (Isso pode ser contrastado com os *erros* u_i, os quais têm n graus de liberdade na amostra.)

Para referência, vamos resumir essa discussão com o Teorema 3.3. Provamos esse teorema para o caso da análise de regressão simples no Capítulo 2 (veja o Teorema 2.3). (No Apêndice E, disponível no site de Thomson, é dada uma prova geral, que requer álgebra matricial.)

TEOREMA 3.3 (ESTIMADOR NÃO-VIESADO DE σ^2)
Sob as hipóteses de Gauss-Markov RLM.1 a RLM.5, $E(\hat{\sigma}^2) = \sigma^2$.

A raiz quadrada positiva de $\hat{\sigma}^2$, denominada $\hat{\sigma}$, é chamada **Erro-Padrão da Regressão (EPR)**. O EPR é um estimador do desvio-padrão do termo erro. Essa estimativa é usualmente informada pelos programas de regressão, embora ela seja chamada de nomes diferentes pelos diferentes programas. (Além de EPR, $\hat{\sigma}$ também é chamado *erro-padrão da estimativa* e a *raiz do erro quadrado médio*.)

Observe que $\hat{\sigma}$ pode diminuir ou aumentar quando outra variável independente é acrescentada a uma regressão (para uma dada amostra). Isso ocorre pois, embora SQR deva cair quando outra variável explicativa é adicionada, os graus de liberdade também diminuem em um. Como SQR está no numerador e gl está no denominador, não podemos dizer, de antemão, qual efeito prevalecerá.

Para construir intervalos de confiança e conduzir testes no Capítulo 4, precisaremos estimar o **desvio-padrão de $\hat{\beta}_j$**, que é exatamente a raiz quadrada da variância:

$$dp(\hat{\beta}_j) = \sigma/[SQT_j(1 - R_j^2)]^{1/2}.$$

Como σ é desconhecido, ele é substituído pelo seu estimador, $\hat{\sigma}$. Isso nos dá o **erro-padrão de $\hat{\beta}_j$**:

$$ep(\hat{\beta}_j) = \hat{\sigma}/[SQT_j(1 - R_j^2)]^{1/2}. \quad (3.58)$$

Assim como as estimativas de MQO podem ser obtidas para qualquer amostra dada, os erros-padrão também podem. Como $ep(\hat{\beta}_j)$ depende de $\hat{\sigma}$, o erro-padrão tem uma distribuição amostral, que desempenhará um papel importante no Capítulo 4.

Devemos enfatizar algo sobre os erros-padrão. Como (3.58) é obtido diretamente da fórmula da variância em (3.51), e como (3.51) se apóia na hipótese de homoscedasticidade RLM.5, a fórmula do erro-padrão em (3.58) *não* é um estimador válido de dp($\hat{\beta}_j$) se os erros exibem heteroscedasticidade. Assim, enquanto a presença de heteroscedasticidade não causa viés em $\hat{\beta}_j$, ela leva viés da fórmula usual da Var($\hat{\beta}_j$), o que invalida, portanto, os erros-padrão. Isso é importante porque qualquer programa de regressão calcula (3.58) como o erro-padrão básico de cada coeficiente (com uma interpretação um pouco diferente para o intercepto). Se suspeitarmos de heteroscedasticidade, então os erros-padrão de MQO "habituais" não são válidos, e alguma ação corretiva deve ser tomada. No Capítulo 8, veremos quais métodos estão disponíveis para trabalhar com a heteroscedasticidade.

3.5 EFICIÊNCIA DE MQO: O TEOREMA DE GAUSS-MARKOV

Nesta seção, apresentaremos e discutiremos o importante **Teorema de Gauss-Markov**, que justifica o uso do método de MQO em vez de usar uma variedade de estimadores concorrentes. Já conhecemos uma justificativa para MQO: sob as hipóteses RLM.1 a RLM.4, MQO é não-viesado. Entretanto, há *muitos* estimadores não-viesados de β_j sob essas hipóteses (por exemplo, veja o Problema 3.12). Poderia haver outros estimadores não-viesados com variâncias menores que as dos estimadores de MQO?

Se limitarmos apropriadamente a classe de estimadores concorrentes, podemos mostrar que MQO *é* o melhor dentro de sua classe. Especificamente, argumentamos que, sob as hipóteses RLM.1 a RLM.5, o estimador de MQO $\hat{\beta}_j$ para β_j é o **melhor estimador linear não-viesado** (*Best Linear Unbiased Estimator — BLUE*). A fim de formular o teorema, precisamos entender cada componente da sigla "BLUE". Primeiro, sabemos o que é um estimador: ele é uma regra que pode ser aplicada a qualquer amostra de dados para produzir uma estimativa. Também sabemos o que é um estimador não-viesado: no contexto corrente, um estimador, por exemplo $\tilde{\beta}_j$, de β_j é um estimador não-viesado de β_j se $E(\tilde{\beta}_j) = \beta_j$ para qualquer $\beta_0, \beta_1, ..., \beta_k$.

E o que dizer sobre o significado do termo "linear"? No contexto atual, um estimador $\tilde{\beta}_j$ de β_j é linear se, e somente se, ele puder ser expresso como uma função linear dos dados da variável dependente:

$$\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i, \qquad \textbf{(3.59)}$$

em que cada w_{ij} pode ser uma função dos valores amostrais de todas as variáveis independentes. Os estimadores de MQO são lineares, como pode ser visto na equação (3.22).

Finalmente, como definir "o melhor"? Para o teorema corrente, o melhor é definido como *a variância menor*. Dados dois estimadores não-viesados, é lógico preferir aquele com a variância menor (veja o Apêndice C, disponível no site de Thomson).

Agora, vamos chamar de $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ os estimadores de MQO do modelo (3.31) sob as hipóteses RLM.1 a RLM.5. O Teorema de Gauss-Markov diz que, para qualquer estimador $\tilde{\beta}_j$ que é *linear* e *não-viesado*, $\text{Var}(\hat{\beta}_j) \leq \text{Var}(\tilde{\beta}_j)$, e a desigualdade é geralmente estrita. Em outras palavras, na classe dos estimadores lineares não-viesados, MQO tem a menor variância (sob as cinco hipóteses de Gauss-Markov). De fato, o teorema diz mais do que isso. Se desejarmos estimar qualquer função linear de β_j, a combinação linear correspondente dos estimadores de MQO alcança a menor variância entre todos os estimadores não-viesados. Vamos concluir com um teorema, o provado no Apêndice 3A, disponível no site da Thomson.

TEOREMA 3.4 (TEOREMA DE GAUSS-MARKOV)

Sob as hipóteses RLM.1 a RLM.5, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ são os melhores estimadores lineares não-viesados (BLUEs) de $\beta_0, \beta_1, \ldots, \beta_k$, respectivamente.

É por causa desse teorema que as hipóteses RLM.1 a RLM.5 são conhecidas como as hipóteses de Gauss-Markov (para a análise de corte transversal).

A importância do Teorema de Gauss-Markov é que, quando um conjunto padrão de hipóteses se mantém, não necessitamos procurar por estimadores não-viesados alternativos da forma expressa em (3.59): nenhum será melhor que MQO. Equivalentemente, se somos apresentados a um estimador que é tanto linear como não-viesado, então sabemos que a variância desse estimador é pelo menos tão grande quanto a variância de MQO; nenhum cálculo adicional é necessário para mostrar isso.

Para nossos propósitos, o Teorema 3.4 justifica o uso de MQO para estimar modelos de regressão múltipla. Se qualquer uma das hipóteses de Gauss-Markov for violada, o teorema não é mais válido. Já sabemos que a falha da hipótese de média condicional zero (hipótese RLM.3) faz com que MQO seja viesado, de modo que o Teorema 3.4 também deixa de ser válido. Também sabemos que a heteroscedasticidade (falha da hipótese RLM.5) não faz com que MQO seja viesado. Entretanto, MQO não tem mais a menor variância entre os estimadores lineares não-viesados na presença da heteroscedasticidade. No Capítulo 8, analisaremos um estimador que melhora MQO quando conhecemos o tipo da heteroscedasticidade.

---

1. O modelo de regressão múltipla nos permite, efetivamente, manter os outros fatores fixos ao examinarmos os efeitos de uma variável independente particular sobre a variável dependente. Ele permite, explicitamente, que as variáveis sejam correlacionadas.

2. Embora o modelo seja linear em seus *parâmetros*, ele pode ser usado para modelar relações não-lineares ao se escolher, apropriadamente, as variáveis dependente e independente.

3. O método de mínimos quadrados ordinários é facilmente aplicado para estimar o modelo de regressão múltipla. Cada estimativa de inclinação mede o efeito parcial da variável independente correspondente sobre a variável dependente, mantendo todas as outras variáveis independentes fixas.

4. R^2 é a proporção da variação amostral da variável dependente explicada pelas variáveis independentes, e é utilizado como uma medida do grau de ajuste. É importante não dar importância demais ao valor do R^2 na avaliação de modelos econométricos.

5. Sob as primeiras quatro hipóteses de Gauss-Markov (RLM.1 a RLM.4), os estimadores de MQO são não-viesados. Isso implica que incluir uma variável irrelevante em um modelo não tem nenhum efeito sobre a inexistência de viés dos estimadores de intercepto e de inclinação. De outro lado, omitir uma variável importante faz com que MQO seja viesado. Em muitas circunstâncias, a direção do viés pode ser determinada.

6. Sob as cinco hipóteses de Gauss-Markov, a variância de um estimador de inclinação de MQO é dada por $\text{Var}(\hat{\beta}_j) = \sigma^2/[\text{SQT}_j(1 - R_j^2)]$. Quando a variância do erro σ^2 cresce, o mesmo ocorre com $\text{Var}(\hat{\beta}_j)$, enquanto $\text{Var}(\hat{\beta}_j)$ diminui quando a variação amostral em x_j, SQT_j, aumenta. O termo R_j^2 mede a magnitude da colinearidade entre x_j e as outras variáveis explicativas. Quando R_j^2 aproximas-se de um, $\text{Var}(\hat{\beta}_j)$ é ilimitada.

7. Adicionar uma variável irrelevante a uma equação geralmente aumenta as variâncias dos demais estimadores de MQO, por causa da multicolinearidade.

8. Sob as hipóteses de Gauss-Markov (RLM.1 a RLM.5), os estimadores de MQO são os melhores estimadores lineares não-viesados (BLUE).

3.1 Usando os dados do arquivo GPA2.RAW sobre 4.137 estudantes de curso superior nos Estados Unidos, estimou-se a seguinte equação por MQO:

$$\widehat{nmgrad} = 1{,}392 - 0{,}0135\ emperc + 0{,}00148\ sat$$
$$n = 4.137,\ R^2 = 0{,}273,$$

em que *nmgrad* é mensurada em uma escala de quatro pontos, *emperc* é o percentil da turma de formados do ensino médio (definido de modo que, por exemplo, *emperc* = 5 significa os cinco por cento melhores da sala), e *sat* é uma nota média ponderada de matemática e habilidade verbal do estudante para ingresso em curso superior.

(i) Por que faz sentido que o coeficiente de *emperc* seja negativo?
(ii) Qual é o valor previsto de *nmgrad* quando *emperc* = 20 e *sat* = 1.050?
(iii) Suponha que dois alunos do ensino médio, A e B, estejam no mesmo percentil no ensino médio, mas a nota *sat* do Estudante A foi 140 pontos maior (cerca de um desvio-padrão na amostra). Qual é a diferença prevista em *nmgrad* para esses dois estudantes? A diferença é grande?
(iv) Mantendo *emperc* fixo, que diferença na nota *sat* levaria a uma diferença prevista de *nmgrad* de 0,50? Comente sua resposta.

3.2 Os dados do arquivo WAGE2.RAW, sobre homens que trabalham, foram utilizados para estimar a seguinte equação:

$$\widehat{educ} = 10{,}36 - 0{,}094\ irms + 0{,}131\ educm + 0{,}210\ educp$$
$$n = 722,\ R^2 = 0{,}214,$$

em que *educ* é anos de escolaridade formal, *irms* é o número de irmãos, *educm* é anos de escolaridade formal da mãe e *educp* é anos de escolaridade formal do pai.

(i) *irms* tem o efeito esperado? Explique. Mantendo *educm* e *educp* fixos, em quanto deveria *irms* aumentar para reduzir os anos previstos da educação formal em um ano? (Uma resposta incompleta é aceitável aqui.)
(ii) Discuta a interpretação do coeficiente de *educm*.
(iii) Suponha que o Homem A não tenha irmãos, e sua mãe e seu pai tenham, cada um, 12 anos de educação formal. Suponha também que o Homem B não tenha irmãos, e sua mãe e seu pai tenham, cada um, 16 anos de educação formal. Qual é a diferença prevista em anos de educação formal entre B e A?

3.3 O modelo seguinte é uma versão simplificada do modelo de regressão múltipla usado por Biddle e Hamermesh (1990) para estudar a escolha entre o tempo gasto dormindo e trabalhando e para observar outros fatores que afetam o sono:

$$dormir = \beta_0 + \beta_1 trabtot + \beta_2 educ + \beta_3 idade + u,$$

em que *dormir* e *trabtot* (trabalho total) são mensurados em minutos por semana e *educ* e *idade* são mensurados em anos. (Veja também o Problema 2.12.)

(i) Se os adultos escolhem entre dormir e trabalhar, qual é sinal de β_1?

(ii) Que sinais você espera que β_2 e β_3 terão?

(iii) Usando os dados do arquivo SLEEP75.RAW, a equação estimada é

$$\widehat{dormir} = 3.638{,}25 - 0{,}148\, trabtot - 11{,}13\, educ + 2{,}20\, idade$$
$$n = 706,\ R^2 = 0{,}113.$$

Se alguém trabalha 5 horas a mais por semana, qual é a queda, em minutos, no valor esperado de dormir? Esse valor representa uma escolha grande?

(iv) Discuta o sinal e a magnitude do coeficiente de *educ*.

(v) Você diria que *trabtot*, *educ* e *idade* explicam muito da variação de *dormir*? Quais outros fatores poderiam afetar o tempo gasto dormindo? É provável que eles sejam correlacionados com *trabtot*?

3.4 O salário inicial (mediano) para recém-formados em direito é determinado pela equação

$$\log(salárioim) = \beta_0 + \beta_1 lsat + \beta_2 nmgrad + \beta_3 \log(volbib) + \beta_4 \log(custo) + \beta_5 rank + u,$$

em que *lsat* é a nota mediana do *lsat* (nota de ingresso no curso de direito) dos recém-formados, *nmgrad* é a nota mediana dos recém-formados nas disciplinas do curso de direito, *volbib* é o número de volumes da biblioteca da escola de direito, *custo* é o custo anual da escola de direito e *rank* é a classificação da escola de direito (com *rank* = 1 sendo o melhor posto de classificação).

(i) Explique a razão de esperarmos $\beta_5 \leq 0$.

(ii) Quais são os sinais que você espera para os outros parâmetros de inclinação? Justifique sua resposta.

(iii) Utilizando os dados do arquivo LAWSCH85.RAW, a equação estimada é

$$\widehat{\log(salárioim)} = 8{,}34 + 0{,}0047\, lsat + 0{,}248\, nmgrad + 0{,}095 \log(volbib)$$
$$+ 0{,}038 \log(custo) - 0{,}0033\, rank$$
$$n = 136,\ R^2 = 0{,}842.$$

Qual é a diferença *ceteris paribus* prevista no salário para as escolas com um *nmgrad* mediano diferente em um ponto? (Descreva sua resposta em percentual.)

(iv) Interprete o coeficiente da variável log(*volbib*).

(v) Você diria que é melhor freqüentar uma escola de direito que tem uma classificação melhor? Qual é a diferença no salário inicial esperado para uma escola que tem uma classificação igual a 20?

3.5 Em um estudo que relaciona a nota média em curso superior (*nmgrad*) ao tempo gasto em várias atividades, você distribui uma pesquisa para vários estudantes. Os estudantes devem responder quan-

tas horas eles despendem, em cada semana, em quatro atividades: estudo, sono, trabalho e lazer. Toda atividade é colocada em uma das quatro categorias, de modo que, para cada estudante, a soma das horas nas quatro atividades deve ser igual a 168.

(i) No modelo

$$nmgrad = \beta_0 + \beta_1 estudar + \beta_2 dormir + \beta_3 trabalhar + \beta_4 lazer + u,$$

faz sentido manter *dormir*, *trabalhar* e *lazer* fixos, enquanto *estudar* varia?

(ii) Explique a razão de esse modelo violar a hipótese RLM.4.

(iii) Como você poderia reformular o modelo, de modo que seus parâmetros tivessem uma interpretação útil e ele satisfizesse a hipótese RLM.4?

3.6 Considere o modelo de regressão múltipla contendo três variáveis independentes, sob as hipóteses RLM.1 a RLM.4:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

Você está interessado em estimar a soma dos parâmetros de x_1 e x_2; chame-a de $\theta_1 = \beta_1 + \beta_2$. Mostre que $\hat{\theta}_1 = \hat{\beta}_1 + \hat{\beta}_2$ é um estimador não-viesado de θ_1.

3.7 Quais dos seguintes itens podem fazer com que os estimadores de MQO sejam viesados?

(i) Heteroscedasticidade.

(ii) Omitir uma variável importante.

(iii) Um coeficiente de correlação amostral de 0,95 entre duas variáveis independentes incluídas no modelo.

3.8 Suponha que a produtividade média do trabalhador da indústria (*prodmed*) dependa de dois fatores — horas médias de treinamento do trabalhador (*treinmed*) e aptidão média do trabalhador (*aptidmed*):

$$prodmed = \beta_0 + \beta_1 treinmed + \beta_2 aptidmed + u.$$

Assuma que essa equação satisfaça as hipóteses de Gauss-Markov. Se um subsídio foi dado às empresas cujos trabalhadores têm uma aptidão menor do que a média, de modo que *treinmed* e *aptidmed* sejam negativamente correlacionados, qual é o provável viés em $\tilde{\beta}_1$ obtido da regressão simples de *prodmed* sobre *treinmed*?

3.9 A equação seguinte descreve o preço mediano das residências de uma comunidade em termos da quantidade de poluição (*oxn*, de óxido nitroso) e do número médio de cômodos nas residências da comunidade (*comods*):

$$\log(preço) = \beta_0 + \beta_1 \log(oxn) + \beta_2 comods + u.$$

(i) Quais são os prováveis sinais de β_1 e β_2? Qual é a interpretação de β_1? Explique.

(ii) Por que *oxn* [ou, mais precisamente, log(*oxn*)] e *comods* deveriam ser negativamente correlacionados? Se esse é o caso, a regressão simples de log(*preço*) sobre log(*oxn*) produz um estimador viesado para cima ou para baixo de β_1?

(iii) Utilizando os dados do arquivo HPRICE2.RAW foram estimadas as seguintes equações:

$$\log(\widehat{preço}) = 11{,}71 - 1{,}043 \log(oxn),\ n = 506,\ R^2 = 0{,}264.$$

$$\log(\widehat{preço}) = 9{,}23 - 0{,}718 \log(oxn) + 0{,}306\ comods,\ n = 506,\ R^2 = 0{,}514.$$

A relação entre as estimativas da elasticidade do *preço* das regressões simples e múltipla é a que você previu, tomando como base suas respostas na parte (ii)? Pode-se dizer que −0,718 está claramente mais próximo da elasticidade verdadeira que −1,043?

3.10 Suponha que o modelo populacional que determina y seja:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

e esse modelo satisfaz as hipóteses de Gauss-Markov. Entretanto, estimamos o modelo que omite x_3. Sejam $\tilde{\beta}_0$, $\tilde{\beta}_1$ e $\tilde{\beta}_2$ os estimadores de MQO da regressão de y sobre x_1 e x_2. Mostre que o valor esperado de $\tilde{\beta}_1$ (dados os valores das variáveis independentes da amostra) é

$$\mathrm{E}(\tilde{\beta}_1) = \beta_1 + \beta_3 \frac{\sum_{i=1}^{n} \hat{r}_{i1} x_{i3}}{\sum_{i=1}^{n} \hat{r}_{i1}^2},$$

em que os $\hat{r}_{i1}$ são os resíduos de MQO da regressão de x_1 sobre x_2. [*Sugestão*: a fórmula de $\tilde{\beta}_1$ é proveniente da equação (3.22). Coloque $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$ nessa equação. Após alguma álgebra, aplique o operador esperança, tratando x_{i3} e $\hat{r}_{i1}$ como não-aleatórios.]

3.11 A seguinte equação representa os efeitos das receitas totais de impostos sobre o crescimento subseqüente do emprego para a população de municípios dos Estados Unidos:

$$cresc = \beta_0 + \beta_1 parc_p + \beta_2 parc_r + \beta_3 parc_v + outros\ fatores,$$

em que *cresc* é a variação percentual do emprego de 1980 a 1990, enquanto o total das receitas de impostos tem a seguinte distribuição: $parc_p$ é a parcela dos impostos sobre a propriedade, $parc_r$ é a parcela das receitas de impostos sobre a renda e $parc_v$ é a parcela das receitas de impostos sobre as vendas. Todas essas variáveis estão mensuradas em 1980. A parcela omitida, $parc_t$, inclui taxas e impostos variados. Por definição, as quatro parcelas somam um. Outros fatores incluiriam despesas com educação, infra-estrutura, e assim por diante (todos mensurados em 1980).

(i) Por que devemos omitir uma das variáveis de parcela de impostos da equação?
(ii) Dê uma interpretação cuidadosa de β_1.

3.12 (i) Considere o modelo de regressão simples $y = \beta_0 + \beta_1 x + u$, sob as primeiras quatro hipóteses de Gauss-Markov. Para alguma função $g(x)$, por exemplo, $g(x) = x^2$ ou $g(x) = \log(1 + x^2)$, defina $z_i = g(x_i)$. Defina um estimador de inclinação como

$$\tilde{\beta}_1 = \left(\sum_{i=1}^{n} (z_i - \bar{z}) y_i\right) \Big/ \left(\sum_{i=1}^{n} (z_i - \bar{z}) x_i\right).$$

Mostre que $\tilde{\beta}_1$ é linear e não-viesado. Lembre-se: como $E(u|x) = 0$, você pode tratar tanto x_i como z_i como não-aleatórios em sua derivação.

(ii) Acrescente a hipótese de homoscedasticidade, RLM.5. Mostre que

$$Var(\tilde{\beta}_1) = \sigma^2\left(\sum_{i=1}^{n}(z_i - \bar{z})^2\right)\Big/\left(\sum_{i=1}^{n}(z_i - \bar{z})\,x_i\right)^2.$$

(iii) Mostre diretamente que, sob as hipóteses de Gauss-Markov, $Var(\hat{\beta}_1) \leq Var(\tilde{\beta}_1)$, em que $\hat{\beta}_1$ é o estimador de MQO. [*Sugestão*: a desigualdade de Cauchy-Schwartz do Apêndice B (disponível no site da Thomson) implica que

$$\left(n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})\right)^2 \leq \left(n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})^2\right)\left(n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right);$$

Observe que podemos retirar $\bar{x}$ da covariância amostral.]

3A.1 Derivação das Condições de Primeira Ordem da Equação (3.13)

A análise é muito similar à do caso da regressão simples. Devemos caracterizar as soluções para o problema

$$\min_{b_0, b_1, \ldots, b_k} \sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - \ldots - b_k x_{ik})^2.$$

Considerando as derivadas parciais em relação a cada um dos b_j (veja o Apêndice A, disponível no site da Thomson), avaliando-as nas soluções e igualando-as a zero resulta

$$-2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0$$

$$-2\sum_{i=1}^{n} x_{ij}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0, \text{ para todo } j = 1, \ldots, k.$$

Cancelando -2 obtemos as condições de primeira ordem em (3.13).

3A.2 Derivação da Equação (3.22)

Para derivar (3.22), escreva x_{i1} em termos de seus valores estimados e seus resíduos a partir da regressão de x_1 sobre $x_2, \ldots, x_k$: $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$, para todo $i = 1, \ldots, n$. Agora, insira essa expressão na segunda equação de (3.13):

$$\sum_{i=1}^{n}(\hat{x}_{i1} + \hat{r}_{i1})(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0 \qquad \textbf{(3.60)}$$

Pela definição do resíduo de MQO $\hat{u}_i$, como $\hat{x}_{i1}$ é exatamente uma função linear das variáveis explicativas x_{i2}, ..., x_{ik}, segue que $\sum_{i=1}^{n} \hat{x}_{i1}\hat{u}_i = 0$. Portanto, a equação (3.60) pode ser expressa como

$$\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}) = 0. \qquad \textbf{(3.61)}$$

Como os $\hat{r}_{i1}$ são os resíduos da regressão de x_1 sobre x_2, ..., x_k, $\sum_{i=1}^{n} x_{ij}\hat{r}_{i1} = 0$, para todo $j = 2, \ldots, k$. Portanto, (3.61) é equivalente a $\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_1 x_{i1}) = 0$. Finalmente, usamos o fato de que $\sum_{i=1}^{n} \hat{x}_{i1}\hat{r}_{i1} = 0$, o que significa que $\hat{\beta}_1$ soluciona

$$\sum_{i=1}^{n} \hat{r}_{i1}(y_i - \hat{\beta}_1 \hat{r}_{i1}) = 0.$$

Agora, com um pouco de álgebra chegamos a (3.22), desde que, evidentemente, $\sum_{i=1}^{n} \hat{r}_{i1}^2 > 0$; isso é garantido pela hipótese RLM.4.

3A.3 Prova do Teorema 3.1

Vamos provar o Teorema 3.1 para $\hat{\beta}_1$; a prova para os outros parâmetros de inclinação é, virtualmente, idêntica. (Veja o Apêndice E, disponível no site da Thomson para uma prova mais sucinta utilizando matrizes.) Sob a hipótese RLM.4, os estimadores de MQO existem, e podemos escrever $\hat{\beta}_1$ como em (3.22). Sob a hipótese RLM.1, podemos escrever y_i como em (3.32); substitua-o pelo y_i de (3.22). Então, usando $\sum_{i=1}^{n} \hat{r}_{i1} = 0$, $\sum_{i=1}^{n} x_{j1}\hat{r}_{i1} = 0$, para todo $j = 2, \ldots, k$ e $\sum_{i=1}^{n} x_{j1}\hat{r}_{i1} = \sum_{i=1}^{n} \hat{r}_{i1}^2$, temos

$$\hat{\beta}_1 = \beta_1 + \left(\sum_{i=1}^{n} \hat{r}_{i1} u_i\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right). \qquad \textbf{(3.62)}$$

Agora, sob as hipóteses RLM.2 e RLM.3, o valor esperado de cada u_i, dadas todas as variáveis independentes na amostra, é zero. Como os $\hat{r}_{i1}$ são justamente funções das variáveis independentes da amostra, segue-se que

$$\mathrm{E}(\hat{\beta}_1 | \boldsymbol{X}) = \beta_1 + \left(\sum_{i=1}^{n} \hat{r}_{i1}\mathrm{E}(u_i|\boldsymbol{X})\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right)$$

$$= \beta_1 + \left(\sum_{i=1}^{n} \hat{r}_{i1} \cdot 0\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right) = \beta_1,$$

em que $\boldsymbol{X}$ representa os dados de todas as variáveis independentes, e $\mathrm{E}(\hat{\beta}_1 | \boldsymbol{X})$ é o valor esperado de $\hat{\beta}_1$, dado x_{i1}, ..., x_{ik}, para todo $i = 1, \ldots, n$. Isso completa a prova.

3A.4 Viés de Variável Omitida no Modelo Geral

Podemos derivar o viés de variável omitida no modelo geral da equação (3.31) sob as quatro primeiras hipóteses de Gauss-Markov. Em particular, sejam $\hat{\beta}_j$, $j = 0, 1, ..., k$ os estimadores de MQO da regressão ao se usar o conjunto completo de variáveis explicativas. Sejam $\tilde{\beta}_j$, $j = 0, 1, ..., k - 1$ os estimadores de MQO da regressão que omite x_k. Sejam $\tilde{\delta}_j$, $j = 1, ..., k - 1$ os coeficientes de inclinação de x_j da regressão auxiliar de x_{ik} sobre $x_{i1}, x_{i2}, ..., x_{i,k-1}$, $i = 1, ..., n$. Um fato útil é que

$$\tilde{\beta}_j = \hat{\beta}_j + \hat{\beta}_k \tilde{\delta}_j. \tag{3.63}$$

Isso mostra explicitamente que, quando não controlamos x_k na regressão, o efeito parcial estimado de x_j é igual ao efeito parcial quando incluímos x_k mais o efeito parcial de x_k sobre $\hat{y}$ vezes a relação parcial entre a variável omitida, x_k, e x_j, $j < k$. Condicionado ao conjunto inteiro de variáveis explicativas, $\boldsymbol{X}$, sabemos que os $\hat{\beta}_j$ são todos não-viesados para os correspondentes β_j, $j = 1, ..., k$. Além disso, como $\tilde{\delta}_j$ é exatamente uma função de $\boldsymbol{X}$, temos

$$\begin{aligned} E(\tilde{\beta}_j|\boldsymbol{X}) &= E(\hat{\beta}_j|\boldsymbol{X}) + E(\hat{\beta}_k|\boldsymbol{X})\,\tilde{\delta}_j \\ &= \beta_j + \beta_k \tilde{\delta}_j. \end{aligned} \tag{3.64}$$

A equação (3.64) mostra que $\tilde{\beta}_j$ é viesado para β_j, a menos que $\beta_k = 0$ — caso em que x_k não tem efeito parcial na população —, ou $\tilde{\delta}_j$ é igual a zero, o que significa que x_{ik} e x_{ij} são parcialmente não-correlacionados na amostra. A chave para obter a equação (3.64) é a equação (3.63). Para mostrar a equação (3.63), podemos usar a equação (3.22) várias vezes. Por simplicidade, vamos olhar para $j = 1$. Agora, $\tilde{\beta}_1$ é o coeficiente de inclinação da regressão simples de y_i sobre $\tilde{r}_{i1}$, $i = 1 ..., n$, em que os $\tilde{r}_{i1}$ são os resíduos de MQO da regressão de x_{i1} sobre $x_{i2}, x_{i3}, ..., x_{i,k-1}$. Considere o numerador da expressão de $\tilde{\beta}_1$: $\sum_{i=1}^{n} \tilde{r}_{i1} y_i$. Para cada i, podemos escrever $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik} + \hat{u}_i$ e colocar em y_i. Agora, pelas propriedades dos resíduos de MQO, os $\tilde{r}_{i1}$ têm média amostral zero e são não-correlacionados com $x_{i2}, x_{i3}, ..., x_{i,k-1}$ na amostra. Semelhantemente, os $\hat{u}_i$ têm média amostral zero e correlação amostral zero com $x_{i1}, x_{i2}, ..., x_{ik}$. Segue-se que os $\tilde{r}_{i1}$ e $\hat{u}_i$ são não-correlacionados na amostra (visto que os $\tilde{r}_{i1}$ são exatamente uma combinação linear de $x_{i1}, x_{i2}, ..., x_{i,k-1}$). Assim

$$\sum_{i=1}^{n} \tilde{r}_{i1} y_i = \hat{\beta}_1 \left(\sum_{i=1}^{n} \tilde{r}_{i1} x_{i1} \right) + \hat{\beta}_k \left(\sum_{i=1}^{n} \tilde{r}_{i1} x_{ik} \right). \tag{3.65}$$

Agora, $\sum_{i=1}^{n} \tilde{r}_{i1} x_{i1} = \sum_{i=1}^{n} \tilde{r}_{i1}^2$, que é também o denominador de $\tilde{\beta}_1$. Portanto, mostramos que

$$\begin{aligned} \tilde{\beta}_1 &= \hat{\beta}_1 + \hat{\beta}_k \left(\sum_{i=1}^{n} \tilde{r}_{i1} x_{ik} \right) \Big/ \left(\sum_{i=1}^{n} \tilde{r}_{i1}^2, \right) \\ &= \hat{\beta}_1 + \hat{\beta}_k \tilde{\delta}_1. \end{aligned}$$

Essa é a relação que queríamos mostrar.

3A.5 Prova do Teorema 3.2

Novamente, vamos provar o teorema para $j = 1$. Escreva $\hat{\beta}_1$ como na equação (3.62). Agora, sob RLM.5, $\text{Var}(u_i|\boldsymbol{X}) = \sigma^2$, para todo $i = 1, ..., n$. Sob amostragem aleatória, os u_i são independentes, mesmo condicionados a $\boldsymbol{X}$, e os $\hat{r}_{i1}$ são não-aleatórios condicionados a $\boldsymbol{X}$. Portanto,

$$\begin{aligned}\text{Var}(\hat{\beta}_1|\boldsymbol{X}) &= \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \,\text{Var}(u_i|\boldsymbol{X})\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right)^2 \\ &= \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \sigma^2\right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right)^2 = \sigma^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2\right).\end{aligned}$$

Agora, visto que $\sum_{i=1}^{n} \hat{r}_{i1}^2$ é a soma dos quadrados dos resíduos da regressão de x_1 sobre $x_2, ..., x_k$, $\sum_{i=1}^{n} \hat{r}_{i1}^2 = \text{SQT}_1(1 - R_1^2)$. Isso completa a prova.

3A.6 Prova do Teorema 3.4

Mostramos que, para qualquer outro estimador linear não-viesado $\tilde{\beta}_1$ de β_1, $\text{Var}(\tilde{\beta}_1) \geq \text{Var}(\hat{\beta}_1)$, em que $\hat{\beta}_1$ é o estimador de MQO. Não se perde generalidade ao jogarmos o foco em $j = 1$.

Para $\tilde{\beta}_1$ como na equação (3.59), podemos inserir em y_i para obter

$$\tilde{\beta}_1 = \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1} x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1} x_{i2} + ... + \beta_k \sum_{i=1}^{n} w_{i1} x_{ik} + \sum_{i=1}^{n} w_{i1} u_i.$$

Agora, visto que os w_{i1} são funções de x_{ij},

$$\begin{aligned}\text{E}(\tilde{\beta}_1|\boldsymbol{X}) &= \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1}x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1}x_{i2} + ... + \beta_k \sum_{i=1}^{n} w_{i1}x_{ik} + \sum_{i=1}^{n} w_{i1}\,\text{E}(u_i|\boldsymbol{X}) \\ &= \beta_0 \sum_{i=1}^{n} w_{i1} + \beta_1 \sum_{i=1}^{n} w_{i1}x_{i1} + \beta_2 \sum_{i=1}^{n} w_{i1}x_{i2} + ... + \beta_k \sum_{i=1}^{n} w_{i1}x_{ik}\end{aligned}$$

porque $\text{E}(u_i|\boldsymbol{X}) = 0$, para todo $i = 1, ..., n$, sob RLM.2 e RLM.3. Portanto, para $\text{E}(\tilde{\beta}_1|\boldsymbol{X})$ igualar-se a β_1 para quaisquer valores dos parâmetros, devemos ter

$$\sum_{i=1}^{n} w_{i1} = 0, \quad \sum_{i=1}^{n} w_{i1} x_{i1} = 1, \quad \sum_{i=1}^{n} w_{i1} x_{ij} = 0, \quad j = 2, ..., k. \tag{3.66}$$

Agora, sejam $\hat{r}_{i1}$ os resíduos da regressão de x_{i1} sobre $x_{i2}, ..., x_{ik}$. Então, de (3.66), segue-se que

$$\sum_{i=1}^{n} w_{i1} \hat{r}_{i1} = 1 \tag{3.67}$$

visto que $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$ e $\sum_{i=1}^{n} w_{i1} \hat{x}_{i1} = 0$. Agora, considere a diferença entre $\text{Var}(\tilde{\beta}_1|\boldsymbol{X})$ e $\text{Var}(\hat{\beta}_1|\boldsymbol{X})$ sob RLM.1 a RLM.5:

$$\sigma^2 \sum_{i=1}^{n} w_{i1}^2 - \sigma^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right). \tag{3.68}$$

Por causa de (3.67), podemos escrever a diferença em (3.68), sem σ^2, como

$$\sum_{i=1}^{n} w_{i1}^2 - \left(\sum_{i=1}^{n} w_{i1} \hat{r}_{i1} \right)^2 \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right). \tag{3.69}$$

No entanto, (3.69) é simplesmente

$$\sum_{i=1}^{n} (w_{i1} - \hat{\gamma}_1 \hat{r}_{i1})^2, \tag{3.70}$$

em que $\hat{\gamma}_1 = \left(\sum_{i=1}^{n} w_{i1} \hat{r}_{i1} \right) \Big/ \left(\sum_{i=1}^{n} \hat{r}_{i1}^2 \right)$, como pode ser visto ao se elevar cada termo em (3.70) ao quadrado, somando e cancelando os termos. Como (3.70) é exatamente a soma dos resíduos quadrados da regressão simples de w_{i1} sobre $\hat{r}_{i1}$ — lembre-se de que a média amostral de $\hat{r}_{i1}$ é zero —, (3.70) deve ser não-negativo. Isso completa a prova.

Análise de Regressão Múltipla: Inferência

Este capítulo continua o estudo da análise de regressão múltipla. Vamos nos voltar agora para o problema de testar hipóteses sobre os parâmetros do modelo da regressão populacional. Iniciaremos encontrando as distribuições dos estimadores de MQO sob a hipótese adicional de que o erro populacional é normalmente distribuído. As seções 4.2 e 4.3 cobrem os testes de hipóteses sobre os parâmetros individuais, enquanto a Seção 4.4 discute como testar uma única hipótese que envolve mais de um parâmetro. Na Seção 4.5, vamos focalizar o teste de restrições múltiplas, bem como dedicar atenção especial em determinar se um grupo de variáveis independentes pode ser omitido do modelo.

4.1 DISTRIBUIÇÕES AMOSTRAIS DOS ESTIMADORES DE MQO

Até este ponto, construímos um conjunto de hipóteses sob as quais o método MQO é não-viesado; também derivamos e discutimos o viés causado por variáveis omitidas. Na Seção 3.4, obtivemos as variâncias dos estimadores de MQO sob as hipóteses de Gauss-Markov. Na seção 3.5, mostramos que essa variância é a menor entre os estimadores lineares não-viesados.

Conhecer o valor esperado e a variância dos estimadores de MQO é útil para descrever sua precisão. Entretanto, para a inferência estatística necessitamos conhecer mais do que apenas os dois primeiros momentos de $\hat{\beta}_j$; precisamos conhecer a distribuição amostral completa de $\hat{\beta}_j$. Mesmo sob as hipóteses de Gauss-Markov, a distribuição de $\hat{\beta}_j$ pode ter, virtualmente, qualquer forma.

Quando estabelecemos um condicionamento aos valores das variáveis independentes de nossa amostra, fica claro que as distribuições amostrais dos estimadores de MQO dependem da distribuição subjacente dos erros. Para tornar as distribuições amostrais de $\hat{\beta}_j$ passíveis de tratamento, vamos assumir agora que o erro não-observado é *normalmente distribuído* na população. Chamamos isso de **hipótese da normalidade**.

HIPÓTESE RLM.6 (NORMALIDADE)

O erro populacional u é independente das variáveis explicativas $x_1, x_2, \ldots, x_k$ e é normalmente distribuído, com média zero e variância σ^2: $u \sim \text{Normal}(0,\sigma^2)$.

A hipótese RLM.6 é muito mais forte que qualquer uma das nossas hipóteses anteriores. De fato, como u é independente de x_j sob RLM.6, $E(u|x_1, ..., x_k) = E(u) = 0$ e $Var(u|x_1, ..., x_k) = Var(u) = \sigma^2$. Assim, ao fazermos a hipótese RLM.6, necessariamente estamos assumindo RLM.3 e RLM.5. Para enfatizar que estamos assumindo mais do que antes, vamos nos referir ao conjunto completo de hipóteses RLM.1 a RLM.6.

Nas aplicações da regressão de corte transversal, as hipóteses RLM.1 a RLM.6 são chamadas **hipóteses do modelo linear clássico (MLC)**. Assim, vamos nos referir ao modelo sob essas seis hipóteses como o **modelo linear clássico**. É melhor pensar as hipóteses do MLC como contendo todas as hipóteses de Gauss-Markov *mais* a hipótese de um termo erro normalmente distribuído.

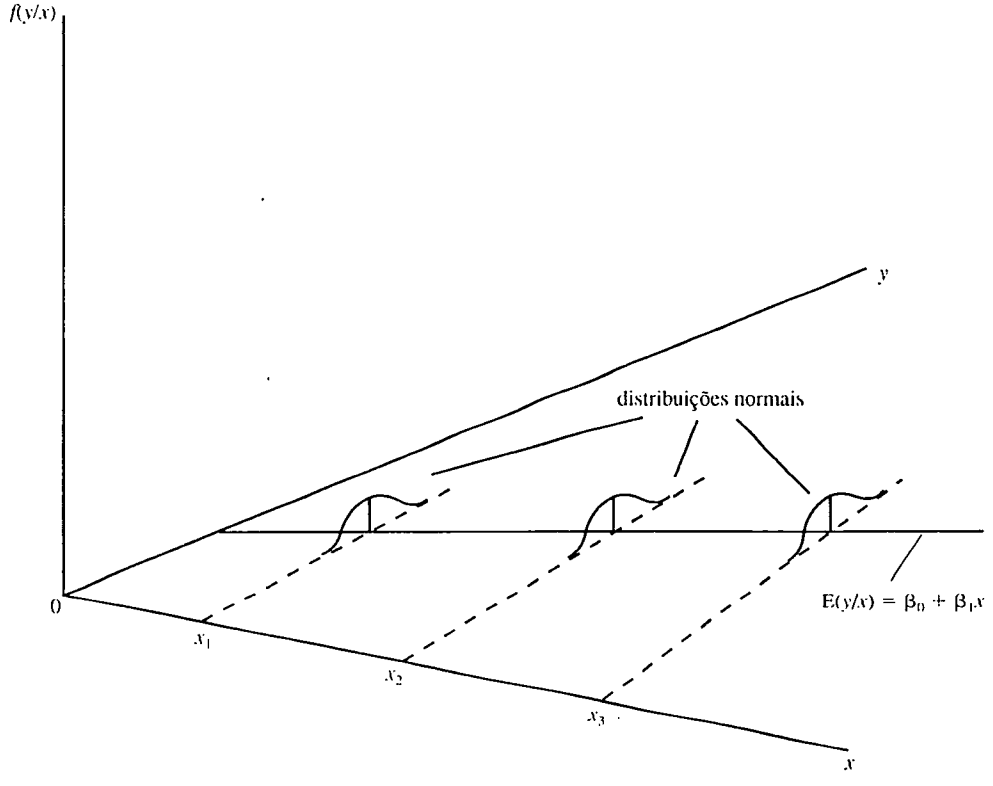
Sob as hipóteses do MLC, os estimadores de MQO $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ têm uma propriedade de eficiência mais forte do que teriam sob as hipóteses de Gauss-Markov. Pode-se mostrar que os estimadores de MQO são os **estimadores não-viesados de variância mínima**, o que significa que MQO tem a menor variância entre os estimadores não-viesados; não temos mais de restringir nossa comparação com os estimadores que são lineares em y_i. Essa propriedade de MQO sob as hipóteses do MLC é discutida mais adiante, no Apêndice E.

Uma maneira sucinta de resumir as hipóteses do MLC na população é

$$y|\boldsymbol{x} \sim \text{Normal}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, \sigma^2),$$

em que $\boldsymbol{x}$ é, novamente, uma maneira de escrever $(x_1, ..., x_k)$. Assim, condicionado a x, y tem uma distribuição normal com média linear em $x_1, ..., x_k$ e uma variância constante. Para uma única variável independente x, essa situação está ilustrada na Figura 4.1.

Figura 4.1
A distribuição normal homoscedástica com uma única variável explicativa.



O argumento para justificar a distribuição normal dos erros é usualmente este: como u é a soma de muitos fatores diferentes não-observados que afetam y, podemos invocar o teorema do limite central (veja Apêndice C, disponível na página do livro, no site www.thomsonlearning.com.br) para concluir que u tem uma distribuição normal aproximada. Esse argumento tem algum mérito, mas não sem debilidades. Primeiro, os fatores em u podem ter distribuições muito diferentes na população (por exemplo, a aptidão e a qualidade da educação no erro de uma equação de salário). Embora o teorema do limite central (TLC) possa ainda ser válido em tais casos, a aproximação à normal pode ser insatisfatória, dependendo de quantos fatores aparecem em u e de quão diferentes são suas distribuições.

Um problema mais sério com o argumento do TLC é que ele assume que todos os fatores não-observados afetam y de um modo separado e aditivo. Nada garante que isso seja assim. Se u é uma função complicada dos fatores não-observados, então o argumento do TLC realmente não se aplica.

Em qualquer aplicação, saber se a normalidade de u pode ser assumida é uma questão empírica. Por exemplo, não há teorema dizendo que *salárioh* condicionado a *educ*, *exper* e *perm* é normalmente distribuído. De qualquer modo, o simples raciocínio sugere que o oposto é verdadeiro: visto que o salário por hora nunca pode ser menor que zero, ele não pode, estritamente falando, ter uma distribuição normal. Além disso, visto que há leis de salário mínimo, alguma fração da população ganha exatamente o salário mínimo, o que viola a hipótese de normalidade. Contudo, como uma questão prática, podemos perguntar se a distribuição condicional do salário está "próxima" de ser normal. A evidência empírica passada sugere que normalidade *não* é uma boa hipótese para os salários.

Freqüentemente, fazer uma transformação, especialmente tomando o log, produz uma distribuição que está mais próxima da normal. Por exemplo, algo como log(*preço*) tende a ter uma distribuição que parece mais normal do que a distribuição de *preço*. Uma vez mais, essa é uma questão empírica. Discutiremos as conseqüências da normalidade para a inferência estatística no Capítulo 5.

Há alguns exemplos em que RLM.6 é claramente falsa. Sempre que y assume apenas alguns valores, ela não pode ter uma distribuição próxima de uma distribuição normal. A variável dependente do Exemplo 3.5 dá um bom exemplo. A variável *npre86*, o número de vezes que um homem jovem foi preso em 1986, assume um conjunto pequeno de valores inteiros e é igual a zero para a maioria dos homens. Assim, *npre86* está longe de ser normalmente distribuída. O que pode ser feito nesses casos? Como veremos no Capítulo 5 – e isso é importante –, a normalidade dos erros não é um problema sério com tamanhos grandes de amostra. Por ora, vamos apenas fazer a hipótese da normalidade.

A normalidade do termo erro traduz-se nas distribuições normais amostrais dos estimadores de MQO:

TEOREMA 4.1 (DISTRIBUIÇÕES AMOSTRAIS NORMAIS)

Sob as hipóteses do MLC, RLM.1 a RLM.6, condicional aos valores amostrais das variáveis independentes,

$$\hat{\beta}_j \sim \text{Normal}[\beta_j, \text{Var}(\hat{\beta}_j)], \tag{4.1}$$

onde $\text{Var}(\hat{\beta}_j)$ foi estudada no Capítulo 3 [equação (3.51)]. Portanto,

$$(\hat{\beta}_j - \beta_j)/\text{dp}(\hat{\beta}_j) \sim \text{Normal}(0,1).$$

A prova de (4.1) não é tão difícil, dadas as propriedades das variáveis aleatórias normalmente distribuídas descritas no Apêndice B, no site da Thomson. Cada $\hat{\beta}_j$ pode ser escrito como $\hat{\beta}_j = \beta_j + \sum_{i=1}^{n} w_{ij} u_i$, onde $w_{ij} = \hat{r}_{ij}/\text{SQR}_j$, $\hat{r}_{ij}$ é o i-ésimo resíduo da regressão de x_j sobre todas as outras variáveis independentes, e SQR_j é a soma dos resíduos quadrados dessa regressão [veja a equação (3.62)]. Como os w_j dependem somente das variáveis independentes, eles podem ser tratados com não-aleatórios. Assim, $\hat{\beta}_j$ é exatamente uma combinação linear dos erros na amostra, $\{u_i: i = 1, 2, ..., n\}$. Sob a hipótese RLM.6 (e a hipótese de amostragem aleatória RLM.2), os erros são variáveis aleatórias independentes e identicamente distribuídas com distribuição Normal $(0, \sigma^2)$. Um fato importante sobre variáveis aleatórias normais independentes é que uma combinação linear de tais variáveis é normalmente distribuída (veja Apêndice B, disponível no site da Thomson). Isso basicamente completa a prova. Na Seção 3.3, vimos que $\text{E}(\hat{\beta}_j) = \beta_j$, e derivamos a $\text{Var}(\beta_j)$ na Seção 3.4; não há necessidade de derivarmos novamente essas expressões.

QUESTÃO 4.3

Suponha que u é independente das variáveis explicativas, e que assume os valores −2, −1, 0, 1 e 2 com probabilidade igual a 1/5. Isto infringe as hipóteses de Gauss-Markov? Isto infringe as hipóteses do MLC?

A segunda parte deste teorema segue imediatamente do fato de que, quando padronizamos uma variável aleatória normal ao subtrair dela sua média e dividi-la pelo seu desvio-padrão, obtemos uma variável aleatória normal padronizada.

As conclusões do Teorema 4.1 podem ser fortalecidas. Além de (4.1), qualquer combinação linear dos $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ também é normalmente distribuída, e qualquer subconjunto dos $\hat{\beta}_j$ tem uma distribuição normal *conjunta*. Esses fatos estão na base dos resultados dos testes no restante deste capítulo. No Capítulo 5 mostraremos que a normalidade dos estimadores de MQO ainda é *aproximadamente* verdadeira em amostras grandes, mesmo sem normalidade dos erros.

4.2 TESTES DE HIPÓTESES SOBRE UM ÚNICO PARÂMETRO POPULACIONAL: O TESTE *t*

Esta Seção cobre o importante tópico de testar hipóteses sobre um único parâmetro da função de regressão populacional. O modelo populacional pode ser escrito como

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u, \qquad (4.2)$$

e assumimos que ele satisfaz as hipóteses do MLC. Sabemos que MQO produz estimadores não-viesados de β_j. Nesta seção, estudaremos como testar hipóteses sobre um particular β_j. Para um entendimento completo dos testes de hipóteses, devemos recordar que os β_j são características desconhecidas da população, e nunca os conheceremos com certeza. No entanto, podemos fazer *hipóteses* sobre o valor de β_j e, em seguida, utilizar inferência estatística para testar nossa hipótese.

A fim de construir os testes de hipóteses, precisamos do seguinte resultado:

TEOREMA 4.2 (A DISTRIBUIÇÃO *t* PARA OS ESTIMADORES PADRONIZADOS)

Sob as hipóteses do MLC, RLM.1 a RLM.6,

$$(\hat{\beta}_j - \beta_j)/\text{ep}(\hat{\beta}_j) \sim t_{n-k-1}, \qquad (4.3)$$

em que $k + 1$ é o número de parâmetros desconhecidos do modelo populacional $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$ (k parâmetros de inclinação mais o intercepto β_0).

Esse resultado difere do Teorema 4.1 em alguns aspectos notáveis. O Teorema 4.1 mostrou que, sob as hipóteses do MLC, $(\hat{\beta}_j - \beta_j)/\text{ep}(\hat{\beta}_j) \sim \text{Normal}(0,1)$. A distribuição t em (4.3) é proveniente do fato de que a constante σ em $\text{ep}(\hat{\beta}_j)$ foi substituída pela variável aleatória $\hat{\sigma}$. A prova de que isso leva a uma distribuição t com $n - k - 1$ graus de liberdade não é particularmente percebida. Essencialmente, a prova mostra que (4.3) pode ser escrita como a razão da variável aleatória normal padronizada $(\hat{\beta}_j - \beta_j)/\text{ep}(\hat{\beta}_j)$ sobre a raiz quadrada de $\hat{\sigma}^2/\sigma^2$. Pode-se mostrar que essas variáveis aleatórias são independentes e $(n - k - 1)\,\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-k-1}$. O resultado decorre da definição de uma variável aleatória t (veja Seção B.5 disponível no *site* da Thomson).

O Teorema 4.2 é importante porque ele nos permite testar hipóteses que envolvem os β_j. Na maioria das aplicações, nosso principal interesse é testar a **hipótese nula**

$$H_0: \beta_j = 0, \qquad (4.4)$$

em que j corresponde a qualquer uma das k variáveis independentes. É importante entender o que (4.4) significa e ser capaz de descrever essas hipóteses em uma linguagem simples em uma determinada aplicação. Como β_j mede o efeito parcial de x_j sobre (o valor esperado de) y, após controlar todas as outras variáveis independentes, (4.4) significa que, uma vez que $x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ foram considerados, x_j não tem *nenhum efeito* sobre o valor esperado de y. Não podemos expressar a hipótese nula como "x_j tem realmente um efeito parcial sobre y" porque isso é verdadeiro para qualquer outro valor de β_j que não zero. O teste clássico é apropriado para testar *hipóteses simples* como (4.4).

Como um exemplo, considere a equação do salário

$$\log(salárioh) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 perm + u.$$

A hipótese nula $H_0: \beta_2 = 0$ significa que, uma vez que a educação formal e a permanência foram consideradas, o número de anos no mercado de trabalho (*exper*) não tem nenhum efeito sobre o salário horário. Essa é uma hipótese economicamente interessante. Se ela é verdadeira, ela implica que o histórico de trabalho de uma pessoa, anterior ao emprego atual, não afeta o salário. Se $\beta_2 > 0$, então a experiência prévia de trabalho contribui para a produtividade e, portanto, para o salário.

Você provavelmente se lembra, de seus cursos de estatística, dos rudimentos do teste de hipótese para a média de uma população normal. (Há uma revisão no Apêndice C, no site da Thomson.) Os mecanismos do teste (4.4) no contexto da regressão múltipla são muitos semelhantes. A parte difícil está em obter as estimativas dos coeficientes, os erros-padrão e os valores críticos, mas a maior parte desse trabalho é feita automaticamente por programas econométricos. Nosso trabalho é aprender como o resultado da regressão pode ser usado para testar as hipóteses de interesse.

A estatística que usamos para testar (4.4) (contra qualquer alternativa) é chamada "a" **estatística *t*** ou "a" **razão *t*** de $\hat{\beta}_j$, e é definida como

$$t_{\hat{\beta}_j} \equiv \hat{\beta}_j / \text{ep}(\hat{\beta}_j). \qquad (4.5)$$

Colocamos "a" entre aspas porque, como veremos em breve, para testar outras hipóteses sobre é necessária uma forma mais geral da estatística *t*. Por ora, é importante saber que (4.5) é apropriada somente para testar (4.4). Quando não houver possibilidade de causar nenhuma confusão, algumas vezes escreveremos *t* no lugar de $t_{\hat{\beta}_j}$.

A estatística *t* de $\hat{\beta}_j$ é simples de calcular, dados $\hat{\beta}_j$ e seu erro-padrão. De fato, a maioria dos programas de regressão faz a divisão automaticamente e informa a estatística *t* juntamente com cada coeficiente e seu erro-padrão.

Antes de discutir como usar (4.5) para testar, formalmente, H_0: $\beta_j = 0$, é útil ver porque $t_{\hat{\beta}_j}$ tem características que o tornam razoável enquanto uma estatística de teste para detectar $\beta_j \neq 0$. Em primeiro lugar, como $\text{ep}(\hat{\beta}_j)$ é sempre positivo, $t_{\hat{\beta}_j}$ tem o mesmo sinal de $\hat{\beta}_j$: se $\hat{\beta}_j$ é positivo, então do mesmo modo é $t_{\hat{\beta}_j}$, e se $\hat{\beta}_j$ é negativo, igualmente negativo é $t_{\hat{\beta}_j}$. Segundo, para um dado valor de $\text{ep}(\hat{\beta}_j)$, um valor maior de $\hat{\beta}_j$ leva a valores maiores de $t_{\hat{\beta}_j}$. Se $\hat{\beta}_j$ fica mais negativo, do mesmo modo fica $t_{\hat{\beta}_j}$.

Como estamos testando H_0: $\beta_j = 0$, é natural olhar nosso estimador não-viesado de β_j, $\hat{\beta}_j$, como um guia. Em qualquer aplicação interessante, a estimativa pontual $\hat{\beta}_j$ *nunca* será exatamente zero, seja H_0 verdadeira ou não. A questão é: Quão distante está $\hat{\beta}_j$ de zero? Um valor amostral de $\hat{\beta}_j$ muito distante de zero fornece evidência contra H_0: $\hat{\beta}_j = 0$. Entretanto, devemos reconhecer que há um erro amostral em nossa estimativa $\hat{\beta}_j$, de modo que o tamanho de $\hat{\beta}_j$ deve ser ponderado pelo seu erro amostral. Como o erro-padrão de $\hat{\beta}_j$ é uma estimativa do desvio-padrão de $\hat{\beta}_j$, $t_{\hat{\beta}_j}$ mede quantos desvios-padrão estimados $\hat{\beta}_j$ estão afastados de zero. Isso é precisamente o que fazemos ao testar se a média de uma população é zero usando a estatística *t* padrão da estatística introdutória. Valores de $t_{\hat{\beta}_j}$ suficientemente distantes de zero resultarão em uma rejeição de H_0. A regra de rejeição depende da hipótese alternativa e do nível de significância escolhido do teste.

Determinar uma regra para rejeitar (4.4) a um dado nível de significância — isto é, a probabilidade de rejeitar H_0 quando ela é verdadeira — requer conhecer a distribuição amostral de $t_{\hat{\beta}_j}$ quando H_0 é verdadeira. Do Teorema 4.2, sabemos que ela é t_{n-k-1}. Esse é o resultado teórico essencial necessário para testar (4.4).

Antes de continuarmos, é importante lembrar que estamos testando hipóteses sobre parâmetros *populacionais*. Não estamos testando hipóteses sobre estimativas de uma amostra particular. Assim, nunca fará sentido formular a hipótese nula como "H_0: $\hat{\beta}_1 = 0$" ou, ainda pior, como "H_0: 0,237 = 0" quando a estimativa do parâmetro for 0,237 na amostra. Estamos testando se o valor populacional desconhecido, β_1, é zero.

Alguns tratamentos da análise de regressão definem a estatística *t* como o *valor absoluto* de (4.5), de modo que a estatística *t* sempre é positiva. Essa prática tem a desvantagem de tornar um pouco confuso o teste contra hipóteses alternativas unilaterais. Ao longo deste livro, a estatística *t* sempre tem o mesmo sinal da estimativa do coeficiente de MQO correspondente.

Teste contra Hipóteses Alternativas Unilaterais

A fim de determinar uma regra para rejeitar H_0, precisamos decidir sobre a **hipótese alternativa** relevante. Primeiro, considere uma **hipótese alternativa unilateral** do tipo

$$H_1: \beta_j > 0. \qquad (4.6)$$

Isso significa que não nos preocupamos com alternativas de H_0 do tipo $H_1: \beta_j < 0$; por alguma razão, talvez tomando como base a introspecção ou a teoria econômica, estamos excluindo os valores populacionais de β_j menores que zero. (Outra maneira de pensar a respeito é que a hipótese nula é realmente $H_0: \beta_j \leq 0$; em qualquer caso, a estatística $t_{\hat{\beta}_j}$ é usada como a estatística de teste.)

Como devemos escolher uma regra de rejeição? Em primeiro lugar, devemos decidir sobre um **nível de significância** ou uma probabilidade de rejeitar H_0 quando ela é, de fato, verdadeira. Em termos mais concretos, suponha que decidimos por um nível de significância de 5%, já que esta é a escolha mais comum. Assim, estamos dispostos a rejeitar erroneamente H_0, quando ela é verdadeira 5% das vezes. Agora, embora $t_{\hat{\beta}_j}$ tenha uma distribuição t sob H_0 — de modo que ele tem média igual a zero —, sob a hipótese alternativa $\beta_j > 0$, o valor esperado de $t_{\hat{\beta}_j}$ é positivo. Assim, estamos procurando um valor positivo "suficientemente grande" de $t_{\hat{\beta}_j}$ a fim de rejeitar $H_0: \beta_j = 0$ em favor de $H_1: \beta_j > 0$. Valores negativos de $t_{\hat{\beta}_j}$ não fornecem evidência em favor de H_1.

A definição de "suficientemente grande", com um nível de significância de 5%, é o 95º percentil de uma distribuição t com $n - k - 1$ graus de liberdade; denominemos esse ponto de c. Em outras palavras, a **regra de rejeição** é que H_0 é rejeitada em favor de H_1, ao nível de significância de 5%, se

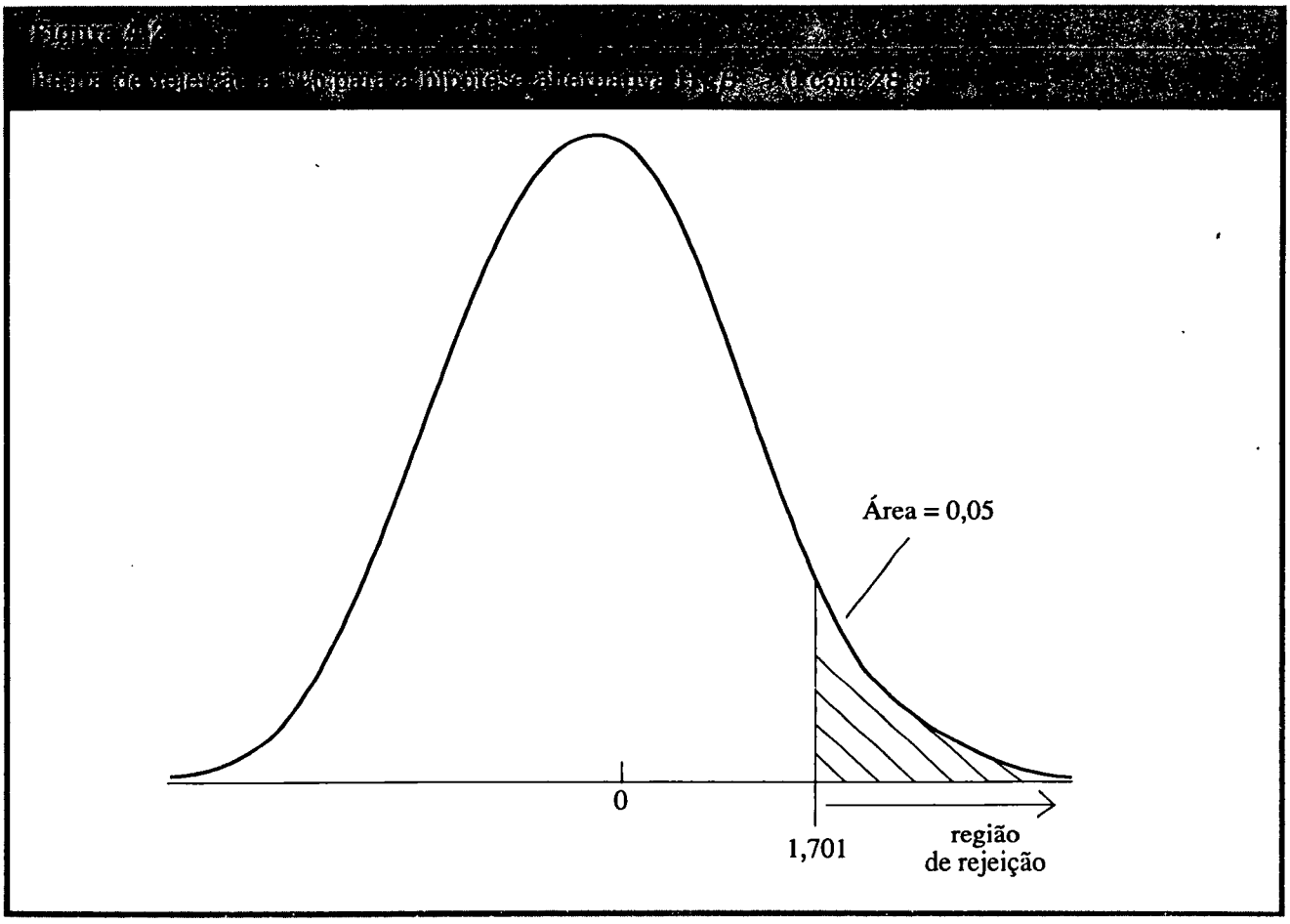
$$t_{\hat{\beta}_j} > c. \qquad (4.7)$$

De acordo com nossa escolha do **valor crítico** c, a rejeição de H_0, quando ela é verdadeira, ocorrerá em 5% de todas as amostras aleatórias.

A regra de rejeição em (4.7) é um exemplo de um **teste monocaudal**. Para obter c, necessitamos somente do nível de significância e dos graus de liberdade. Por exemplo, para um teste ao nível de 5% e com $n - k - 1 = 28$ graus de liberdade, o valor crítico é $c = 1{,}701$. Se $t_{\hat{\beta}_j} < 1{,}701$, não rejeitamos H_0 em favor de (4.6) ao nível de 5%. Observe que um valor negativo de $t_{\hat{\beta}_j}$, não importando o tamanho desse valor em termos absolutos, leva a uma negativa em rejeitar H_0 em favor de (4.6). (Veja a Figura 4.2.)

O mesmo procedimento pode ser usado com outros níveis de significância. Para um teste ao nível de 10% e se $gl = 21$, o valor crítico é $c = 1{,}323$. Para um nível de significância de 1% e se $gl = 21$, $c = 2{,}518$. Todos esses valores críticos são obtidos diretamente da Tabela G.2, do Apêndice G. Você deve observar um padrão nos valores críticos: quando o nível de significância cai, o valor crítico aumenta, de modo que, para rejeitar H_0, exigimos um valor cada vez maior de $t_{\hat{\beta}_j}$. Assim, se H_0 é rejeitada, por exemplo, ao nível de 5%, então ela também é, automaticamente, rejeitada ao nível de 10%. Não faz sentido rejeitar a hipótese nula, por exemplo, ao nível de 5% e, em seguida, refazer o teste para nos certificarmos do resultado ao nível de 10%.

Quando os graus de liberdade da distribuição t ficam maiores, a distribuição t aproxima-se da distribuição normal padronizada. Por exemplo, quando $n - k - 1 = 120$, o valor crítico de 5% para a hipótese alternativa unilateral (4.7) é 1,658, comparável ao valor normal padronizado de 1,645. Esses valores, para objetivos práticos, são suficientemente próximos; para graus de liberdade maiores que 120, pode-se usar os valores críticos da distribuição normal padronizada.



EXEMPLO 4.1

(Equação do Salário Horário)

Usando os dados do arquivo WAGE1.RAW, obtemos a equação estimada

$$\log(salárioh) = 0{,}284 + 0{,}092\, educ + 0{,}0041\, exper + 0{,}022\, perm$$
$$(0{,}104)\quad (0{,}007)\qquad (0{,}0017)\qquad (0{,}003)$$
$$n = 526,\ R^2 = 0{,}316,$$

em que os erros-padrão aparecem em parênteses abaixo dos coeficientes estimados. Seguiremos essa convenção ao longo do livro. Essa equação pode ser usada para testar se o retorno de *exper*, controlando *educ* e *perm*, é zero na população, contra a hipótese alternativa de que ele é positivo. Para tanto, escreva H_0: $\beta_{exper} = 0$ *versus* H_1: $\beta_{exper} > 0$. (Nas aplicações, indexar um parâmetro pelo nome da variável ao qual está

EXEMPLO 4.1 (continuação)

associado é uma maneira hábil de caracterizar os parâmetros, visto que os índices numéricos que usamos são, em geral, arbitrários e podem causar confusão.) Lembre-se de que β_{exper} representa o parâmetro populacional desconhecido. Não faz sentido algum escrever "H_0: 0,0041 = 0" ou "H_0: $\hat{\beta}_{exper} = 0$".

Como temos 522 graus de liberdade, podemos usar os valores críticos da distribuição normal padronizada. O valor crítico a 5% é 1,645, e o valor crítico a 1% é 2,326. A estatística *t* para $\hat{\beta}_{exper}$ é

$$t_{\hat{\beta}_{exper}} = 0{,}0041/0{,}0017 \approx 2{,}41,$$

e portanto $\hat{\beta}_{exper}$, ou *exper*, é estatisticamente significante mesmo ao nível de 1%. Também dizemos que "$\hat{\beta}_{exper}$ é estatisticamente maior que zero ao nível de significância de 1%".

O retorno estimado para um ano a mais de experiência, mantendo fixas a permanência e a educação formal, não é muito grande. Por exemplo, acrescentar três anos a mais aumenta log(*salárioh*) em 3(0,0041) = 0,0123, de modo que o salário é somente cerca de 1,2% maior. No entanto, mostramos, de modo convincente, que o efeito parcial da experiência é positivo na população.

A hipótese alternativa unilateral cujo parâmetro é menor que zero,

$$H_1: \beta_j < 0, \tag{4.8}$$

também aparece nas aplicações. A regra de rejeição para a hipótese alternativa (4.8) é exatamente a imagem espelhada do caso anterior. Agora, o valor crítico vem da cauda esquerda da distribuição *t*. Na prática, é mais fácil pensar a regra de rejeição como

$$t_{\hat{\beta}_j} < -c, \tag{4.9}$$

em que *c* é o valor crítico da hipótese alternativa H_1: $\beta_j > 0$. Para simplificar, sempre assumimos que *c* é positivo, visto que é assim que os valores críticos são apresentados nas tabelas *t*, e, portanto, o valor crítico $-c$ é um número negativo.

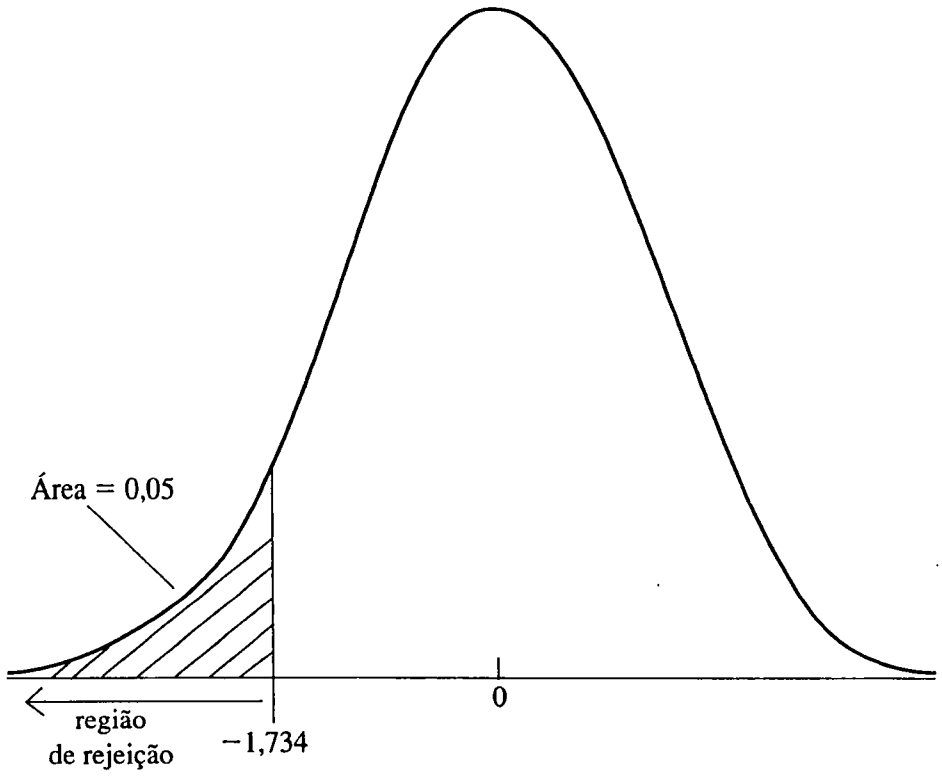
Sejam as taxas de aprovação de empréstimos de uma cidade determinadas por

$$taxaprov = \beta_0 + \beta_1\, porcmin + \beta_2\, rendmed + \beta_3 riquemed + \beta_4\, dívidamed + u,$$

em que *porcmin* é a percentagem de minorias na cidade, *rendmed* é a renda média, *riquemed* é a riqueza média e *dívidamed* é alguma medida de dívidas médias. Como você formularia a hipótese nula de que *não* há diferença nas taxas de empréstimos entre os bairros devido à composição racial e étnica quando a renda média, a riqueza média e a dívida média foram controladas? Como você formularia a hipótese alternativa de que há discriminação contra as minorias nas taxas de aprovação de empréstimos?

Por exemplo, se o nível de significância é 5% e o número de graus de liberdade é 18, então $c = 1{,}734$, e assim H_0: $\beta_j = 0$ é rejeitada em favor de H_1: $\beta_j < 0$ ao nível de 5% se $t_{\hat{\beta}_j} < -1{,}734$. É importante lembrar que, para rejeitar H_0 contra a hipótese alternativa negativa (4.8), devemos obter uma estatística t negativa. Uma razão t positiva, não importa o quão grande ela seja, não fornece evidência em favor de (4.8). A regra de rejeição está ilustrada na Figura 4.3.

Figura 4.3
Regra de rejeição a 5% para a hipótese alternativa H_1: $\beta_j < 0$, com 18 *gl*.



EXEMPLO 4.2

(Desempenho de Estudantes e Tamanho de Escolas)

Há muito interesse no efeito do tamanho das escolas sobre o desempenho dos estudantes. (Veja, por exemplo, *The New York Times Magazine*, 28.5.95.) Afirma-se que, tudo o mais sendo igual, os estudantes de escolas menores saem-se melhor do que aqueles de escolas maiores. Assume-se que essa hipótese é verdadeira, mesmo após considerar as diferenças nos tamanhos das salas entre as escolas.

O arquivo MEAP93.RAW contém dados sobre 408 escolas de ensino médio em Michigan para o ano de 1993. Podemos usar esses dados para testar a hipótese nula de que o tamanho da escola não tem efeito sobre as notas de testes padronizados, contra a hipótese alternativa de que o tamanho tem um efeito negativo. O desempenho é medido pela percentagem de estudantes que recebem uma nota de aprovação no teste de matemática, *mate10*. O tamanho da escola é medido pelo número de estudantes matriculados (*matricl*). A hipótese nula é H_0: $\beta_{matricl} = 0$, e a hipótese alternativa é H_1: $\beta_{matricl} < 0$. Por ora, vamos controlar outros

EXEMPLO 4.2 (continuação)

dois fatores: o salário anual médio dos professores (*totsal*) e o número de funcionários por mil estudantes (*staff*). O salário do professor é uma medida da qualidade do professor, e o tamanho de *staff* é uma medida aproximada de quanta atenção os estudantes recebem.

A equação estimada, com os erros-padrão entre parênteses, é

$$\widehat{maîe10} = 2{,}274 + 0{,}00046\ totsal + 0{,}048\ staff - 0{,}00020\ matricl$$
$$(6{,}113)\quad (0{,}00010)\qquad (0{,}040)\qquad (0{,}00022)$$
$$n = 408,\ R^2 = 0{,}0541.$$

O coeficiente de *matricl*, −0,00020, está de acordo com a conjectura de que escolas maiores tolhem o desempenho: números maiores de matrículas levam a uma percentagem menor de estudantes com uma nota de aprovação. (Os coeficientes de *totsal* e *staff* também têm os sinais que esperamos.) O fato de *matricl* ter um coeficiente estimado diferente de zero pode ser devido, justamente, ao erro de amostragem; para nos convencermos de um efeito, precisamos conduzir um teste *t*.

Como n − k − 1 = 408 − 4 = 404, usamos o valor crítico normal padronizado. Ao nível de 5%, o valor crítico é −1,65; para rejeitar H_0 ao nível de 5%, a estatística *t* de *matricl* deve ser menor que −1,65.

A estatística de *matricl* é −0,00020/0,00022 ≈ −0,91, que é maior que −1,65: não podemos rejeitar H_0 em favor de H_1 ao nível de 5%. De fato, o valor crítico ao nível de 15% é −1,04, e, como −0,91 > −1,04, não é possível rejeitar H_0 mesmo ao nível de 15%. Concluímos que *matricl* não é estatisticamente significante ao nível de 15%.

A variável *totsal* é estatisticamente significante mesmo ao nível de significância de 1% porque sua estatística *t* é 4,6. Do outro lado, a estatística *t* de *staff* é 1,2, portanto não podemos rejeitar H_0: $\beta_{staff} = 0$ contra H_1: $\beta_{staff} > 0$, mesmo ao nível de significância de 10%. (O valor crítico, calculado a partir da distribuição normal padronizada, é c = 1,28.)

Para ilustrar como a mudança da forma funcional pode afetar nossas conclusões, vamos estimar também o modelo com todas as variáveis independentes na forma logarítmica. Isso permite, por exemplo, que o efeito do tamanho da escola diminua quando o tamanho da escola aumenta. A equação estimada é

$$\widehat{maîe10} = -207{,}66 + 21{,}16\log(totsal) + 3{,}98\log(staff) - 1{,}29\log(matricl)$$
$$(48{,}70)\qquad (4{,}06)\qquad\quad (4{,}19)\qquad\quad (0{,}69)$$
$$n = 408,\ R^2 = 0{,}0654$$

A estatística t de log(*matricl*) é cerca de −1,87; como esse valor está abaixo do valor crítico ao nível de 5%, −1,65, rejeitamos H_0: $\beta_{\log(matricl)} = 0$ em favor de H_1: $\beta_{\log(matricl)} < 0$ ao nível de 5%.

No Capítulo 2, encontramos um modelo em que a variável dependente aparecia em sua forma original (chamada forma em *nível*), enquanto a variável independente aparecia na forma log (chamado modelo *nível-log*). A interpretação dos parâmetros é a mesma no contexto da regressão múltipla, exceto, evidentemente, que podemos dar uma interpretação *ceteris paribus* aos parâmetros. Mantendo fixos *totsal* e *staff*, temos $\Delta\widehat{maîe10} = -1{,}29[\Delta\log(matricl)]$, de modo que

$$\Delta\widehat{maîe10} \approx -(1{,}29/100)(\%\Delta matricl) \approx -0{,}013(\%\Delta matricl).$$

EXEMPLO 4.2 (continuação)

Uma vez mais, usamos o fato de que a variação em log(*matricl*), quando multiplicada por 100, é aproximadamente a variação percentual em *matricl*. Assim, se o número de matrículas é 10% maior em uma escola, prevê-se que *mate10* é 0,013(10) = 0,13 ponto percentual menor (*mate10* é mensurado como uma percentagem).

Que modelo preferimos: aquele que usa o nível de *matricl* ou aquele que usa log(*matricl*)? No modelo nível-nível, o número de matrículas não tem um efeito estatisticamente significante, mas no modelo nível-log ele tem. Isso se traduz em um R-quadrado maior para o modelo nível-log, o que significa dizer que explicamos mais da variação em *mate10* ao usar *matricl* na forma logarítmica (6,5% contra 5,4%). O modelo nível-log é preferível, pois ele captura, de modo mais próximo, a relação entre *mate10* e *matricl*. No Capítulo 6 falaremos mais sobre como usar o R-quadrado para escolher a forma funcional.

Teste contra Hipóteses Alternativas Bilaterais

Nas aplicações, é comum testar a hipótese nula H_0: $\beta_j = 0$ contra uma hipótese **alternativa bilateral**, ou seja,

$$H_1: \beta_j \neq 0. \qquad \textbf{(4.10)}$$

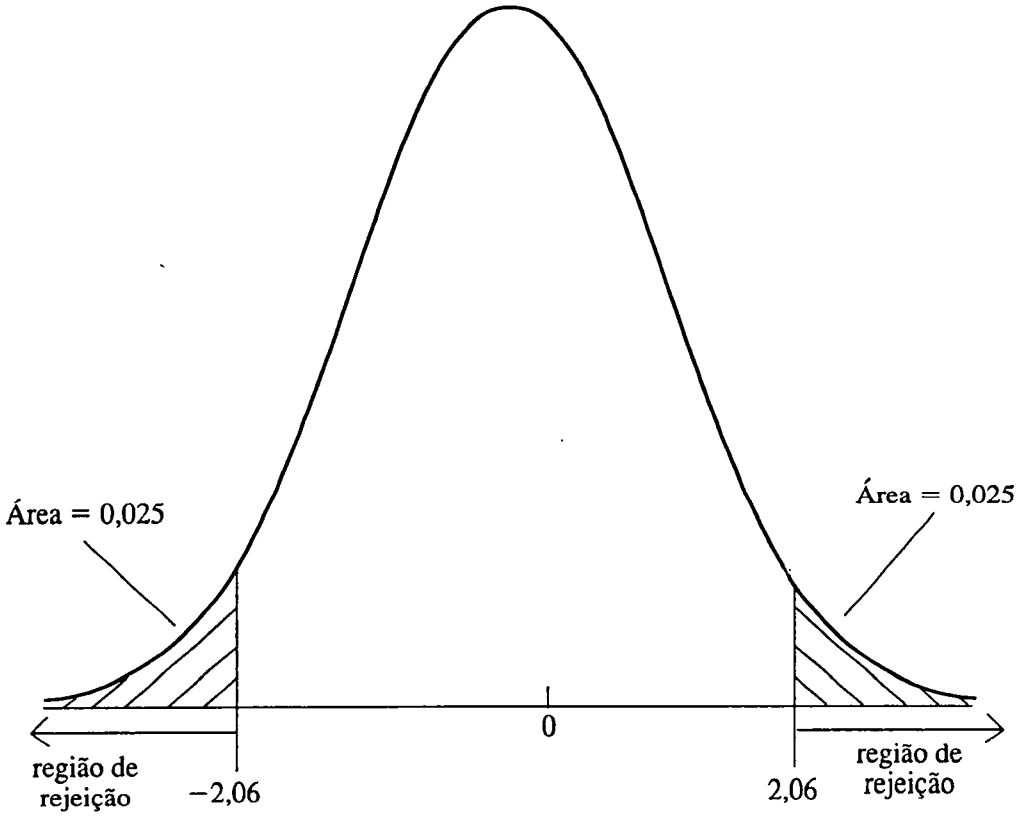
Sob essa hipótese alternativa, x_j tem um efeito *ceteris paribus* sobre y sem especificar se o efeito é positivo ou negativo. Ela é a hipótese alternativa relevante quando o sinal de β_j não é bem-determinado pela teoria (ou pelo senso comum). Mesmo quando sabemos se β_j é positivo ou negativo sob a hipótese alternativa, um teste bilateral é muitas vezes prudente. No mínimo, usar uma alternativa bilateral nos impede de olhar a equação estimada e, então, basear a hipótese alternativa em se $\hat{\beta}_j$ é positivo ou negativo. Usar as estimativas da regressão para nos ajudar a formular as hipóteses nula e alternativa não é permitido, porque a inferência estatística clássica pressupõe que formulamos as hipóteses nula e alternativa sobre a população antes de olhar os dados. Por exemplo, não devemos estimar em primeiro lugar a equação que relaciona o desempenho em matemática ao número de matrículas, observar que o efeito estimado é negativo e, em seguida, decidir que a hipótese alternativa relevante é H_1: $\beta_{matricl} < 0$.

Quando a alternativa é bilateral, estamos interessados no *valor absoluto* da estatística t. A regra de rejeição para H_0: $\beta_j = 0$ contra (4.10) é

$$|t_{\hat{\beta}_j}| > c, \qquad \textbf{(4.11)}$$

em que $|\cdot|$ representa o valor absoluto e c é um valor crítico apropriadamente escolhido. Para achar c, vamos especificar novamente um nível de significância, por exemplo, de 5%. Para um teste **bi-caudal**, c é escolhido de tal forma a fazer com que a área em cada cauda da distribuição t seja igual a 2,5%. Em outras palavras, c é o 97,5º percentil da distribuição t com $n - k - 1$ graus de liberdade. Quando $n - k - 1 = 25$, o valor crítico de 5% para um teste bilateral é $c = 2{,}060$. A Figura 4.4 ilustra essa distribuição.

Figura 4.4
Regra de rejeição a 5% para a hipótese alternativa $H_1: \beta_j \neq 0$ com 25 *gl*.



Quando uma hipótese alternativa específica não é formulada, considera-se geralmente que ela é bilateral. No restante deste livro, o padrão será uma hipótese alternativa bilateral, e 5% será o nível de significância padrão. Ao conduzir uma análise econométrica empírica, é sempre uma boa idéia explicitar qual é a hipótese alternativa e o nível de significância. Se H_0 é rejeitada em favor de (4.10) ao nível de 5%, em geral dizemos que "x_j é **estatisticamente significante**, ou estatisticamente diferente de zero, ao nível de 5%". Se H_0 não é rejeitada, dizemos que "x_j é **estatisticamente não significante** ao nível de 5%".

EXEMPLO 4.3

(Determinantes de *nmgrad*)

Usamos os dados do arquivo GPA1.RAW para estimar um modelo que explique a nota média em curso superior (*nmgrad*), utilizando o número de faltas às aulas por semana (*faltas*) como uma variável explicativa adicional. O modelo estimado é

$$\widehat{nmgrad} = 1{,}39 + 0{,}412\, nmem + 0{,}015\, tac - 0{,}083\, faltas$$
$$(0{,}33)\quad (0{,}094)\qquad (0{,}011)\qquad (0{,}026)$$
$$n = 141,\ R^2 = 0{,}234.$$

EXEMPLO 4.3 (continuação)

Podemos facilmente calcular as estatísticas *t* para verificar quais variáveis são estatisticamente significantes ao usar uma hipótese alternativa bilateral em cada caso. O valor crítico de 5% é cerca de 1,96, visto que os graus de liberdade (141 − 4 = 137) são suficientemente grandes para usar a aproximação normal padronizada. O valor crítico de 1% é cerca de 2,58.

A estatística *t* de *nmem* é 4,38, significante a níveis de significância muito pequenos. Assim, dizemos que "*nmem* é estatisticamente significante a qualquer nível de significância *convencional*". A estatística *t* de *tac* é 1,36, que não é estatisticamente significante ao nível de 10% contra uma alternativa bilateral. O coeficiente de *tac* também é, na prática, pequeno: um aumento de 10 pontos em *tac*, que é grande, faz com que o valor previsto de *nmgrad* cresça somente 0,15 ponto. Assim, a variável *tac* é, na prática, bem como estatisticamente, não significante.

O coeficiente de *faltas* tem uma estatística *t* de −0,083/0,026 = −3,19, de modo que *faltas* é estatisticamente significante ao nível de significância de 1% (3,19 > 2,58). Isso significa que uma falta a mais por semana diminui o *nmgrad* previsto em cerca de 0,083. Assim, mantendo *nmem* e *tac* fixos, a diferença prevista em *nmgrad* entre um estudante que não falta a nenhuma aula por semana e um estudante que falta a cinco aulas por semana é de 0,42. Lembre-se de que isso não diz nada sobre estudantes específicos; referindo-se apenas aos estudantes médios dentro da população.

No exemplo, para cada variável do modelo, poderíamos argumentar que uma hipótese alternativa unilateral é apropriada. As variáveis *nmem* e *faltas* são muito significantes ao se usar um teste bicaudal e têm os sinais que esperamos, de modo que não há razão para fazer um teste monocaudal. Do outro lado, contra uma hipótese alternativa unilateral ($\beta_3 > 0$), *tac* é significante ao nível de 10% mas não ao nível de 5%. Isso não muda o fato de o coeficiente de *tac* ser muito pequeno.

Testes de outras Hipóteses sobre β_j

Embora H_0:$\beta_j = 0$ seja a hipótese mais comum, algumas vezes queremos testar se β_j é igual a alguma outra constante dada. Dois exemplos comuns são $\beta_j = 1$ e $\beta_j = -1$. Em geral, se a hipótese nula é expressa como

$$H_0: \beta_j = a_j, \tag{4.12}$$

em que a_j é o nosso valor hipotético de β_j, então a estatística *t* apropriada é

$$t = (\hat{\beta}_j - a_j)/\text{ep}\,(\hat{\beta}_j).$$

Assim como antes, *t* mede quantos desvios-padrão estimados $\hat{\beta}_j$ está distante do valor hipotético de β_j. A estatística *t* geral é usualmente escrita como

$$t = \frac{(\textit{estimativa} - \textit{valor hipotético})}{\textit{erro-padrão}}. \tag{4.13}$$

Sob (4.12), essa estatística *t* é distribuída como t_{n-k-1}, de acordo com o Teorema 4.2. A estatística *t* usual é obtida quando $a_j = 0$.

Podemos usar a estatística t geral para fazer o teste contra hipóteses alternativas unilaterais ou bilaterais. Por exemplo, se as hipóteses nula e alternativa são H_0: $\beta_j = 1$ e H_1: $\beta_j > 1$, então encontramos o valor crítico para uma alternativa unilateral *exatamente* como antes: a diferença está em como calculamos a estatística t, não em como obtemos o c apropriado. Rejeitamos H_0 em favor de H_1 se $t > c$. Nesse caso, diríamos que "$\hat{\beta}_j$ é estatisticamente maior que um" ao nível de significância apropriado.

EXEMPLO 4.4

(Crimes no *Campus* e Matrículas)

Considere um modelo simples que relaciona o número anual de crimes no *campus* de uma universidade (*crime*) ao número de estudantes matriculados na universidade (*matricl*):

$$\log(crime) = \beta_0 + \beta_1 \log(matricl) + u.$$

Esse é um modelo de elasticidade constante, em que β_1 é a elasticidade do crime em relação às matrículas. Não é muito útil testar H_0: $\beta_1 = 0$, se esperamos que o número total de crimes aumente quando o tamanho do *campus* aumenta. Uma hipótese mais interessante seria supor que a elasticidade do crime em relação a matrículas é igual a um: H_0: $\beta_1 = 1$. Isso significa que um aumento de 1% nas matrículas leva, em média, a um aumento de 1% nos crimes. Uma hipótese alternativa digna de nota é H_1: $\beta_1 > 1$, implicando que um aumento de 1% nas matrículas aumenta o crime no campus em *mais* de 1%. Se $\beta_1 > 1$, então, em um sentido relativo — não exatamente um sentido absoluto —, o crime é mais um problema de *campi* maiores. Uma maneira de ver isso é considerar o exponencial da equação:

$$crime = \exp(\beta_0) matricl^{\beta_1} \exp(u).$$

(Veja o Apêndice A, disponível no site da Thomson, para as propriedades do logaritmo natural e das funções exponenciais.) Para $\beta_0 = 0$ e $u = 0$, essa equação está representada na Figura 4.5, com $\beta_1 < 1$, $\beta_1 = 1$ e $\beta_1 > 1$.

Vamos testar $\beta_1 = 1$ contra $\beta_1 > 1$, usando os dados de 97 faculdades e universidades dos Estados Unidos no ano de 1992, os quais estão contidos no arquivo CAMPUS.RAW. Os dados são provenientes da publicação *FBI's Uniform Crime Reports*, e o número médio de crimes no *campus* é cerca de 394 na amostra, enquanto o número médio de matrículas é aproximadamente 16.076. A equação estimada (com as estimativas e os erros-padrão arredondados em duas casas decimais) é

$$\widehat{\log(crime)} = -6{,}63 + 1{,}27 \log(matricl) \quad \textbf{(4.14)}$$
$$(1{,}03) \quad (0{,}11)$$
$$n = 97, R^2 = 0{,}585.$$

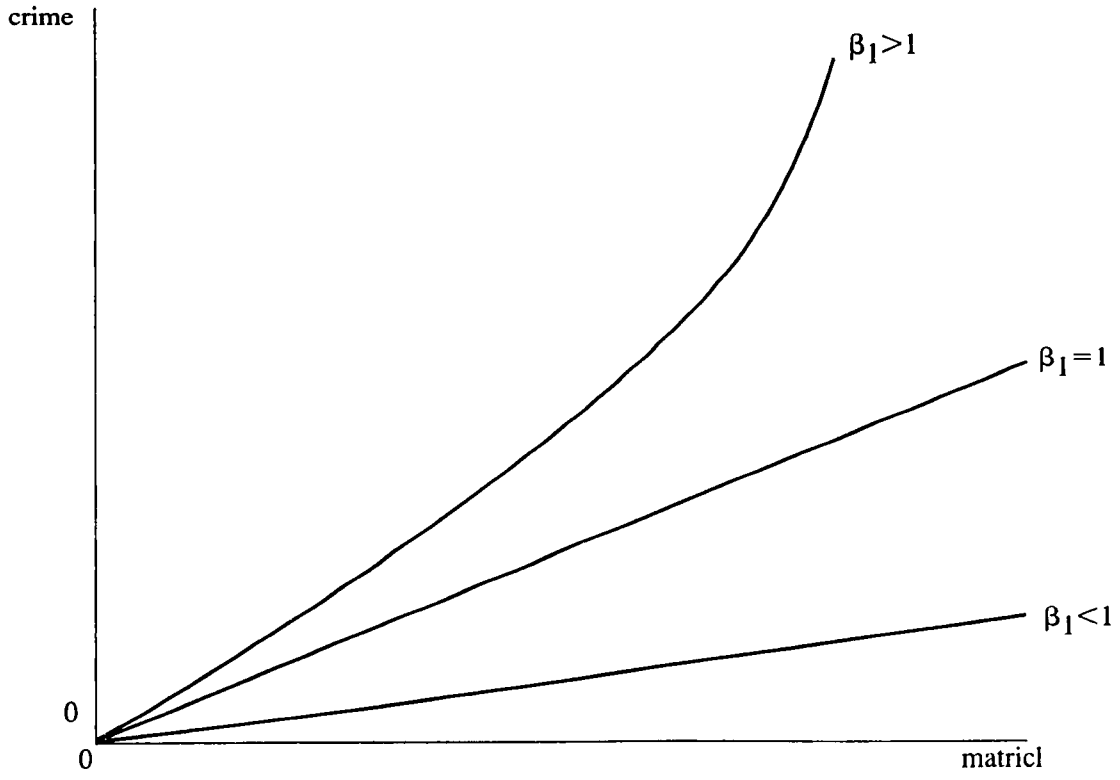
A elasticidade estimada de crime em relação a *matricl*, 1,27, está na direção da hipótese alternativa $\beta_1 > 1$. Porém, há evidência suficiente para concluir que $\beta_1 > 1$? Precisamos tomar cuidado ao testar essa hipótese, especialmente porque os resultados estatísticos dos programas padrão de regressão são muito mais complexos do que o resultado simplificado informado pela equação (4.14). Nosso primeiro instinto deveria ser construir "a" estatística t, tomando o coeficiente de log(*matricl*) e dividindo-o pelo seu erro-padrão, que é a estatística t descrita por um programa de regressão. No entanto, essa é a estatística *errada*

EXEMPLO 4.4 (continuação)

para testar H_0: $\beta_1 = 1$. A estatística t correta é obtida de (4.13): subtraímos o valor hipotético, um, da estimativa e dividimos o resultado pelo seu erro-padrão de $\hat{\beta}_1$: $t = (1{,}27 - 1)/0{,}11 = 0{,}27/0{,}11 \approx 2{,}45$. O valor crítico unilateral de 5% para uma distribuição t com $97 - 2 = 95$ *gl* é cerca de 1,66 (usando *gl* = 120), de modo que, claramente, rejeitamos $\beta_1 = 1$ em favor de $\beta_1 > 1$ ao nível de 5%. De fato, o valor crítico de 1% é cerca de 2,37, e portanto rejeitamos a hipótese nula em favor da hipótese alternativa, ao nível de 1%.

Devemos ter em mente que essa análise não mantém os outros fatores constantes e, portanto, a elasticidade de 1,27 não é necessariamente uma boa estimativa do efeito *ceteris paribus*. É possível que um número maior de matrículas esteja correlacionado com outros fatores que tornam o crime maior: escolas maiores podem estar localizadas em áreas de incidência maior de crimes. Poderíamos controlar isso ao coletar dados sobre taxas de crimes em cada cidade.

Figura 4.5
Gráfico de $crime = matricl^{\beta_1}$ para $\beta_1 < 1$, $\beta_1 = 1$ e $\beta_1 > 1$.



Para um teste alternativo bilateral, por exemplo, H_0: $\beta_j = -1$, H_1: $\beta_j \neq -1$, ainda calculamos a estatística t como em (4.13): $t = (\hat{\beta}_j + 1)/\text{ep}(\hat{\beta}_j)$ (observe que subtrair -1 significa adicionar 1). A regra de rejeição para o teste bicaudal é a usual: rejeitar H_0 se $|t| > c$, em que c é o valor crítico bicaudal. Se H_0 é rejeitada, dizemos que "$\hat{\beta}_j$ é estatisticamente diferente do valor negativo um" ao nível de significância apropriado.

EXEMPLO 4.5

(Preços de Casas e Poluição do Ar)

Para uma amostra de 506 comunidades na área de Boston, estimamos um modelo que relaciona o preço mediano das casas (*preço*) nas comunidades a várias características das comunidades: *oxn* é a quantidade de óxido nitroso no ar, em partes por milhão; *dist* é uma distância ponderada da comunidade em relação a cinco centros de emprego, em milhas; *comods* é o número médio de cômodos nas casas da comunidade; e *razestud* é a razão média estudante-professor nas escolas da comunidade. O modelo populacional é

$$\log(preço) = \beta_0 + \beta_1 \log(oxn) + \beta_2 \log(dist) + \beta_3 comods + \beta_4 razestud + u.$$

Assim, β_1 é a elasticidade do preço em relação a *oxn*. Queremos testar $H_0: \beta_1 = -1$ contra a hipótese alternativa $H_1: \beta_1 \neq -1$. A estatística t para fazer esse teste é $t = (\hat{\beta}_1 + 1)/ep(\hat{\beta}_1)$.

Usando os dados do arquivo HPRICE2.RAW, o modelo estimado é

$$\widehat{\log(preço)} = 11{,}08 - 0{,}954 \log(oxn) - 0{,}134 \log(dist) + 0{,}255\, comods - 0{,}052\, razestud$$

$$(0{,}32) \quad (0{,}117) \quad (0{,}043) \quad (0{,}019) \quad (0{,}006)$$

$$n = 506,\ R^2 = 0{,}581.$$

Todas as estimativas de inclinação têm os sinais esperados. Cada coeficiente é estatisticamente diferente de zero a níveis de significância muito pequenos, incluindo o coeficiente de log(*oxn*). No entanto, não queremos testar $\beta_1 = 0$. A hipótese nula de interesse é $H_0: \beta_1 = -1$, com a estatística t correspondente $(-0{,}954 + 1)/0{,}117 = 0{,}393$. Quando a estatística t é pequena como essa, há pouca necessidade de olhar a tabela t de um valor crítico: a elasticidade estimada não é estatisticamente diferente de -1, mesmo a níveis de significância bastante altos. Controlando fatores que incluímos, há pouca evidência de que a elasticidade seja diferente de -1.

Cálculos dos *p*-Valores dos Testes *t*

Até agora, falamos sobre como testar hipóteses ao usar uma abordagem clássica: após formular a hipótese alternativa, escolhemos um nível de significância, que então determina um valor crítico. Uma vez que o valor crítico tenha sido identificado, o valor da estatística t é comparado com o valor crítico, e a hipótese nula é rejeitada ou não, ao nível de significância dado.

Mesmo após decidir sobre a alternativa apropriada, há um componente de arbitrariedade na abordagem clássica, resultante da necessidade de escolher um nível de significância com antecedência. Diferentes pesquisadores preferem níveis de significância diferentes, dependendo da aplicação particular. Não há nível de significância "correto".

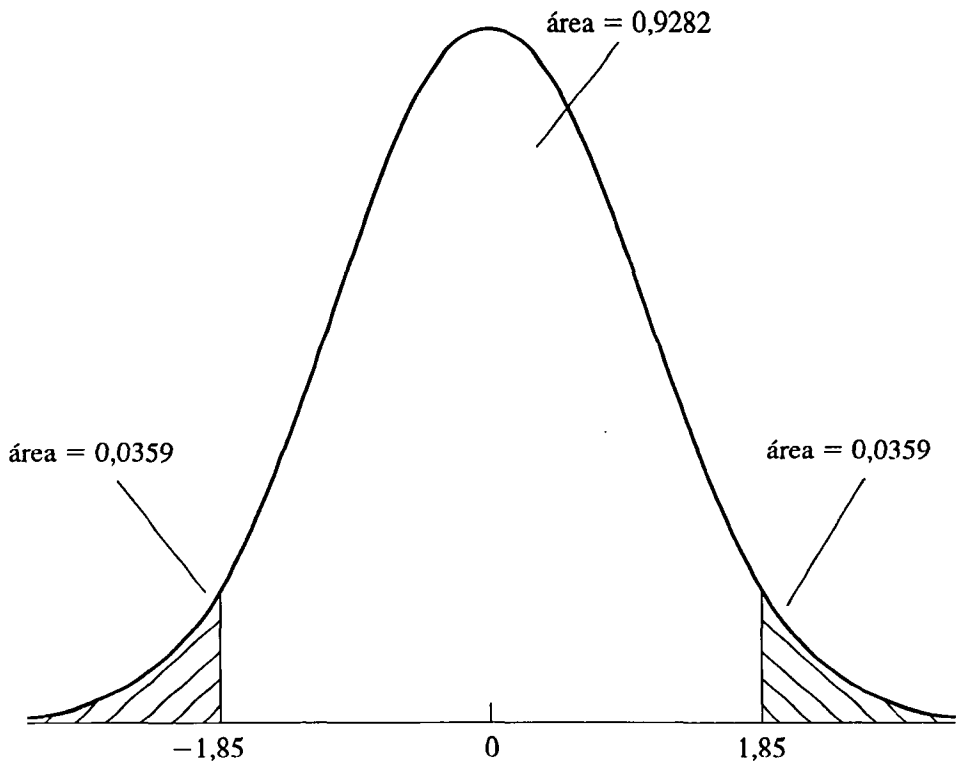
Comprometer-se com um nível de significância antecipadamente pode esconder informações úteis sobre o resultado de um teste de hipóteses. Por exemplo, suponha que desejamos testar a hipótese nula de que um parâmetro seja zero contra uma hipótese alternativa bilateral, e que com 40 graus de liberdade, obtivemos uma estatística t igual a 1,85. A hipótese nula não é rejeitada ao nível de 5%, visto que a estatística t é menor que o valor crítico bicaudal de $c = 2{,}021$. Um pesquisador cujo propósito é não rejeitar a hipótese nula poderia simplesmente descrever esse resultado juntamente com a estimativa: a hipótese nula não é rejeitada ao nível de 5%. Evidentemente, se a estatística t, o coeficiente e seu erro-

padrão fossem informados, então poderíamos também determinar que a hipótese nula seria rejeitada ao nível de 10%, já que o valor crítico de 10% é $c = 1{,}684$.

Em vez de fazer o teste a diferentes níveis de significância, é mais informativo responder à seguinte questão: dado o valor observado da estatística t, qual é *o menor* nível de significância ao qual a hipótese nula seria rejeitada? Esse nível é conhecido como o **p-valor** do teste (veja o Apêndice C, disponível no site da Thomson). No exemplo anterior, sabemos que o p-valor é maior que 0,05, visto que a hipótese nula não é rejeitada ao nível de 5%, e sabemos que o p-valor é menor que 0,10, já que a hipótese nula é rejeitada ao nível de 10%. Obtemos o p-valor real ao calcular a probabilidade de que uma variável aleatória t, com 40 gl, seja maior que 1,85 em valor absoluto. Isto é, o p-valor é o nível de significância do teste quando usamos o valor da estatística de teste (1,85 no exemplo anterior) como o valor crítico do teste. Esse p-valor é mostrado na Figura 4.6.

Figura 4.6

Obtendo o p-valor contra uma alternativa bilateral, quando $t = 1{,}85$ e $gl = 40$.



Como um p-valor é uma probabilidade, seu valor está sempre entre zero e um. A fim de calcular os p-valores, precisamos de tabelas impressas extremamente detalhadas da distribuição t — o que não é muito prático — ou um programa de computador que calcule as áreas sob a função densidade de probabilidade da distribuição t. A maioria dos programas de regressão modernos tem essa capacidade. Alguns deles calculam os p-valores rotineiramente a cada regressão de MQO, mas somente para certas hipóteses. Se um programa de regressão informa um p-valor juntamente com o resultado padrão de MQO, esse valor é, quase certamente, o p-valor de testar a hipótese nula H_0: $\beta_j = 0$ contra a hipótese alternativa bilateral. O p-valor, nesse caso, é

$$P(|T| > |t|), \qquad (4.15)$$

em que, por clareza, T representa uma variável aleatória com distribuição t, com $n - k - 1$ graus de liberdade, e t é o valor numérico da estatística de teste.

O p-valor resume, com precisão, a força e a fraqueza da evidência empírica contra a hipótese nula. Talvez a interpretação mais útil seja a seguinte: o p-valor é a probabilidade de observar uma estatística t tão extrema quanto aceitaríamos *se a hipótese nula fosse verdadeira.* Isso significa que p-valores *pequenos* são evidências *contra* a hipótese nula; p-valores grandes fornecem pouca evidência contra H_0. Por exemplo, se p-valor = 0,50 (informado sempre como um decimal, e não uma percentagem), observaríamos um valor da estatística t tão extremo quanto o faríamos em 50% de todas as amostras aleatórias quando a hipótese nula fosse verdadeira; essa é uma evidência bastante fraca contra H_0.

No exemplo com $gl = 40$ e $t = 1,85$, o p-valor é calculado como

$$p\text{-valor} = P(|T| > 1,85) = 2P(T > 1,85) = 2(0,0359) = 0,0718,$$

em que $P(T > 1,85)$ é a área à direita de 1,85 da distribuição t com 40 gl. (Esse valor foi calculado usando o programa econométrico Stata; ele não está disponível na Tabela G.2.) Isso significa que, se a hipótese nula for verdadeira, observaríamos um valor absoluto da estatística t tão grande quanto 1,85 em cerca de 7,2% das vezes. Isso fornece alguma evidência contra a hipótese nula, mas não a rejeitaríamos ao nível de significância de 5%.

O exemplo anterior ilustra que, uma vez que o p-valor foi calculado, um teste clássico pode ser realizado a qualquer nível desejado. Se α é o nível de significância do teste (na forma decimal), então H_0 é rejeitada se p-valor $< \alpha$; de outro modo, H_0 não é rejeitada ao nível de $100 \cdot \alpha\%$.

Calcular os p-valores de alternativas unilaterais também é muito simples. Suponha, por exemplo, que vamos testar H_0: $\beta_j = 0$ contra H_1: $\beta_j > 0$. Se $\hat{\beta}_j < 0$, então calcular um p-valor não é importante: sabemos que o p-valor é maior que 0,50, o que nunca nos fará rejeitar H_0 em favor de H_1. Se $\hat{\beta}_j > 0$, então $t > 0$ e o p-valor é exatamente a probabilidade de uma variável aleatória, com os gl apropriados, exceder o valor t. Alguns programas de regressão somente calculam os p-valores para alternativas bilaterais. No entanto, é simples obter o p-valor unilateral: apenas divida o p-valor bilateral por 2.

Se a hipótese alternativa for H_1: $\beta_j < 0$, faz sentido calcular um p-valor se $\hat{\beta}_j < 0$ (e portanto $t < 0$): p-valor = $P(T < t) = P(T > |t|)$, porque a distribuição t é simétrica em torno de zero. Uma vez mais, isso pode ser obtido como a metade do p-valor de um teste bicaudal.

Suponha que você tenha estimou um modelo de regressão e obteve $\hat{\beta}_1 = 0,56$ e p-valor = 0,086 para testar H_0: $\beta_1 = 0$ contra H_1: $\beta_1 \neq 0$. Qual é o *p-valor* para testar H_0: $\beta_1 = 0$ contra H_1: $\beta_1 > 0$?

Como as magnitudes das estatísticas t que levam à significância estatística se tornarão rapidamente familiares, especialmente para tamanhos de amostras grandes, não é sempre crucial descrever os p-valores das estatísticas t. No entanto, não é incorreto informá-las. Além disso, quando discutirmos o teste F na Seção 4.5, veremos que é importante calcular os p-valores, porque os valores críticos dos testes F não são facilmente memorizados.

Lembrete sobre a Linguagem do Teste de Hipóteses Clássico

Quando H_0 não é rejeitada, preferimos usar a linguagem "não é possível rejeitar H_0 ao nível de x%" em vez de "H_0 é aceita ao nível de x%". Podemos usar o Exemplo 4.5 para ilustrar o porquê de a primeira afirmação ser preferida. Naquele exemplo, a elasticidade estimada do *preço* em relação a *oxn* é −0,954, e a estatística *t* para testar H_0: $\beta_{oxn} = -1$ é $t = 0{,}393$; portanto, não podemos rejeitar H_0. Porém, há muitos outros valores de β_{oxn} (mais do que podemos contar) que não podem ser rejeitados. Por exemplo, a estatística *t* para H_0: $\beta_{oxn} = -0{,}9$ é $(-0{,}954 + 0{,}9)/0{,}117 = -0{,}462$, e portanto essa hipótese nula também não é rejeitada. Claramente, $\beta_{oxn} = -1$ e $\beta_{oxn} = -0{,}9$ não podem ser ambas verdadeiras, de modo que não faz sentido dizer que "aceitamos" uma dessas hipóteses. Tudo o que podemos dizer é que os dados não nos permitem rejeitar uma dessas hipóteses ao nível de significância de 5%.

Significância Econômica ou Prática *versus* Significância Estatística

Após termos enfatizado a *significância estatística* ao longo desta seção, agora é um bom momento para lembrar que devemos prestar atenção na magnitude das estimativas dos *coeficientes*, além do tamanho das estatísticas *t*. A significância estatística de uma variável x_j é determinada completamente pelo tamanho de $t_{\hat{\beta}_j}$, enquanto a **significância econômica** ou a **significância prática** da variável está relacionada ao tamanho (e sinal) de $\hat{\beta}_j$.

Lembre-se de que a estatística *t* para testar H_0: $\beta_j = 0$ é definida ao dividirmos a estimativa por seu erro-padrão: $t_{\hat{\beta}_j} = \hat{\beta}_j/\text{ep}(\hat{\beta}_j)$. Assim, $t_{\hat{\beta}_j}$ pode indicar significância estatística porque $\hat{\beta}_j$ é "grande" ou porque $\text{ep}(\hat{\beta}_j)$ é "pequeno". É importante, na prática, distinguir entre essas duas razões das estatísticas *t* estatisticamente significantes. Colocar muita ênfase sobre a significância estatística pode levar à conclusão falsa de que uma variável é "importante" para explicar *y* embora seu efeito estimado seja moderado.

EXEMPLO 4.6

[Taxas de Participação nos Planos de Pensão]

No Exemplo 3.3, usamos os dados dos planos de pensão para estimar um modelo que descreve as taxas de participação em termos de taxas de complementação das empresas e das idades dos planos. Vamos incluir, agora, uma medida de tamanho das empresas, o número total de empregados das empresas, (*totemp*). A equação estimada é

$$\underset{}{taxap} = \underset{(0{,}78)}{80{,}29} + \underset{(0{,}52)}{5{,}44}\, taxicont + \underset{(0{,}045)}{0{,}269}\, idade + \underset{(0{,}00004)}{0{,}00013}\, totemp$$

$$n = 1{,}534,\ R^2 = 0{,}100.$$

A menor estatística *t*, em valor absoluto, é a da variável *totemp*: $t = -0{,}00013/0{,}00004 = -3{,}25$, e ela é estatisticamente significante a níveis de significância muito pequenos. (O *p*-valor bicaudal dessa estatística *t* é cerca de 0,001.) Assim, todas as variáveis são estatisticamente significantes a níveis de significância bem pequenos.

Qual o tamanho, em um sentido prático, do coeficiente de *totemp*? Mantendo fixos *taxcomp* e *idade*, se uma firma cresce em 10.000 empregados, a taxa de participação cai em $10.000(0{,}00013) = 1{,}3$ pontos percentuais. Isso é um crescimento enorme no número de empregados, com um efeito somente modesto na taxa de participação. Assim, embora o tamanho da firma afete, de fato, a taxa de participação, o efeito não é, na prática, muito grande.

O exemplo anterior mostra que é particularmente importante interpretar a magnitude do coeficiente, além de olhar as estatísticas *t*, ao trabalhar com amostras grandes. Com tamanhos de amostras grandes, os parâmetros podem ser estimados com muita precisão: os erros-padrão são, em geral, muito pequenos em relação às estimativas dos coeficientes, o que freqüentemente resulta em significância estatística.

Alguns pesquisadores insistem em usar níveis de significância pequenos quando o tamanho da amostra cresce, em parte como uma maneira de compensar o fato de que os erros-padrão estão ficando menores. Por exemplo, se nos sentimos confortáveis com um nível de 5% quando n corresponde a algumas centenas, de observações, deveríamos usar o nível de 1% quando n corresponde a alguns milhares. Usar um nível de significância menor significa que as significâncias econômica e estatística são mais prováveis de coincidir, mas não há garantias: no exemplo anterior, mesmo se usarmos um nível de significância tão pequeno quanto 0,1% (um décimo de um por cento), ainda concluiríamos que *totemp* é estatisticamente significante.

A maior parte dos pesquisadores também está disposta a considerar níveis de significância maiores em aplicações com tamanhos de amostra pequenos, refletindo o fato de que é difícil achar significância com tamanhos de amostra menores (os valores críticos são maiores em magnitude, e os estimadores são menos precisos). Infelizmente, se esse é ou não o caso pode depender dos planos subjacentes do pesquisador.

EXEMPLO 4.7

(Efeitos das Subvenções a Treinamento de Trabalho sobre as Taxas de Rejeição de Produtos das Empresas)

A taxa de rejeição de produtos de uma firma manufatureira é o número de itens defeituosos que devem ser descartados de cada 100 itens produzidos. Assim, uma diminuição nessa taxa de rejeição reflete maior produtividade.

Podemos usar a taxa de rejeição para mensurar o efeito do treinamento dos trabalhadores sobre a produtividade. Para uma amostra de firmas manufatureiras de Michigan em 1987, estimou-se a seguinte equação:

$$\log(rejei) = \underset{(4,91)}{13,72} - \underset{(0,019)}{0,028}\, hrsemp - \underset{(0,41)}{1,21}\log(vendas) + \underset{(0,43)}{1,48}\log(empreg)$$
$$n = 30,\ R^2 = 0,431.$$

(Essa regressão usa um subconjunto de dados de JTRAIN.RAW.) A variável *hrsemp* corresponde às horas anuais de treinamento por trabalhador, *vendas* corresponde às vendas anuais da firma (em dólares), e *empreg* é o número de empregados da firma. A taxa média de rejeição na amostra é cerca de 3,5, e *hrsemp* médio é cerca de 7,3.

A principal variável de interesse é *hrsemp*. Uma hora a mais de treinamento por trabalhador diminui log(*rejei*) em 0,028, o que significa que a taxa de rejeição é cerca de 2,8% menor. Assim, se *hrsemp* aumenta em 5 — cada empregado é treinado 5 horas a mais por ano —, estima-se que a taxa de rejeição caia em 5(2,8) = 14%. Isso parece ser um efeito razoavelmente grande mas, saber se o treinamento adicional vale a pena para a firma, depende do custo de treinamento e dos benefícios de uma taxa de rejeição menor. Não temos os números necessários para fazer uma análise custo/benefício, mas o efeito estimado não parece trivial.

E o que dizer sobre a *significância estatística* da variável de treinamento? A estatística *t* de *hrsemp* é −0,028/0,019 = −1,47, e agora você provavelmente a reconhece como não sendo suficientemente grande em magnitude para concluir que *hrsemp* é estatisticamente significante ao nível de 5%. De fato, com 30 − 4 = 26 graus de liberdade para a alternativa unilateral, $H_1: \beta_{hrsemp} < 0$, o valor crítico de 5% é cerca de −1,71.

EXEMPLO 4.7 (continuação)

Assim, usando um teste de nível estrito a 5%, devemos concluir que *hrsemp* não é estatisticamente significante, mesmo usando uma alternativa unilateral.

Como o tamanho da amostra é bastante pequeno, poderíamos ser mais liberais com o nível de significância. O valor crítico de 10% é $-1,32$, e portanto *hrsemp* é significante contra a alternativa unilateral ao nível de 10%. O p-valor é facilmente calculado como $P(T_{26} < -1,47) = 0,077$. Esse pode ser um p-valor suficientemente pequeno para concluir que o efeito estimado do treinamento não se deve apenas ao erro de amostragem, mas alguns economistas teriam opiniões diferentes a respeito do assunto.

Lembre-se de que erros-padrão grandes podem também ser um resultado da multicolinearidade (alta correlação entre algumas das variáveis independentes), mesmo que o tamanho da amostra pareça razoavelmente grande. Como discutimos na Seção 3.4, não há muito que possamos fazer sobre esse problema além de coletar mais dados ou mudar o escopo da análise excluindo certas variáveis independentes do modelo. Como no caso de um tamanho de amostra pequeno, pode ser difícil estimar precisamente os efeitos parciais quando algumas das variáveis explicativas são altamente correlacionadas. (A Seção 4.5 contém um exemplo.)

Finalizamos esta seção com algumas instruções para discutir as significâncias econômica e estatística de uma variável em um modelo de regressão múltipla:

1. Cheque a significância estatística. Se a variável é estatisticamente significante, discuta a magnitude do coeficiente para ter uma idéia de sua importância prática ou econômica. Esse último passo pode requerer algum cuidado, dependendo de como as variáveis independentes e dependentes aparecem na equação. (Em particular, quais são as unidades de medida? As variáveis aparecem na forma logarítmica?)
2. Se uma variável não é estatisticamente significante aos níveis usuais (10%, 5% ou 1%), você poderia ainda perguntar se a variável tem o efeito esperado sobre y e se tal efeito é, na prática, grande. Se ele é grande, você deve calcular um p-valor para a estatística t. Para tamanhos de amostras pequenos, você pode, às vezes, construir um argumento para p-valores tão grandes quanto 0,20 (mas não há regras rigorosas). Com p-valores grandes, isto é, estatísticas t pequenas, estamos pisando em gelo fino, porque as estimativas grandes, na prática, podem ser devidas ao erro de amostragem: uma amostra aleatória diferente poderia resultar em uma estimativa muito diferente.
3. É comum encontrar variáveis com estatísticas t pequenas que têm o sinal "errado". Para propósitos práticos, elas podem ser ignoradas: concluímos que as variáveis são estatisticamente não significantes. Uma variável importante que tem sinal não esperado e um efeito prático grande é um problema muito mais preocupante e difícil de resolver. Em geral, deve-se pensar mais sobre o modelo e a natureza dos dados, a fim de solucionar tais problemas. Freqüentemente, uma estimativa contra-intuitiva e significante resulta da omissão de uma variável fundamental ou de um dos problemas importantes que discutiremos nos Capítulos 9 e 15.

4.3 INTERVALOS DE CONFIANÇA

Sob as hipóteses do modelo linear clássico, podemos facilmente construir um **intervalo de confiança** (**IC**) para o parâmetro populacional β_j. Os intervalos de confiança são também chamados *estimativas de intervalo*, porque eles dão uma extensão dos valores prováveis do parâmetro populacional, e não somente uma estimativa pontual.

Usando o fato de que $(\hat{\beta}_j - \beta_j)/\text{ep}(\hat{\beta}_j)$ tem uma distribuição t com $n - k - 1$ graus de liberdade [veja (4.3)], uma simples manipulação algébrica leva a um IC do β_j desconhecido. Um *intervalo de confiança de 95%*, é dado por

$$\hat{\beta}_j \pm c \cdot \text{ep}(\hat{\beta}_j), \tag{4.16}$$

em que a constante c é o 97,5º percentil de uma distribuição t_{n-k-1}. Mais precisamente, os limites inferiores e superiores do intervalo de confiança são dados por

$$\underline{\beta}_j \equiv \hat{\beta}_j - c \cdot \text{ep}(\hat{\beta}_j)$$

e

$$\bar{\beta}_j \equiv \hat{\beta}_j - c \cdot \text{ep}(\hat{\beta}_j),$$

respectivamente.

Neste ponto, é útil rever o significado de um intervalo de significância. Se as amostras aleatórias fossem obtidas repetidas vezes, com $\underline{\beta}_j$ e $\bar{\beta}_j$ calculados a cada vez, então o valor populacional (desconhecido) β_j estaria dentro do intervalo $(\underline{\beta}_j, \bar{\beta}_j)$ em 95% das amostras. Infelizmente, para a única amostra que usamos para construir o IC, não sabemos se β_j está, realmente, contido no intervalo. Esperamos que tenhamos obtido uma amostra que seja uma das 95% de todas as amostras em que a estimativa de intervalo contém β_j, mas não temos essa garantia.

Construir um intervalo de confiança é muito simples quando se usa a tecnologia computacional atual. São necessárias três quantidades: $\hat{\beta}_j$, $\text{ep}(\hat{\beta}_j)$ e c. A estimativa do coeficiente e seu erro-padrão são informados por qualquer programa de regressão. Para obter o valor de c, devemos conhecer os graus de liberdade, $n - k - 1$, e o nível de confiança — 95% neste caso. Portanto, o valor de c é obtido da distribuição t_{n-k-1}.

Como um exemplo, para $gl = n - k - 1 = 25$, um intervalo de confiança de 95% para qualquer β_j é dado por $[\hat{\beta}_j - 2{,}06 \cdot \text{ep}(\hat{\beta}_j), \hat{\beta}_j + 2{,}06 \cdot \text{ep}(\hat{\beta}_j)]$.

Quando $n - k - 1 > 120$, a distribuição t_{n-k-1} está suficientemente próxima da normal para usar o 97,5º percentil de uma distribuição normal padrão para construir um IC de 95%: $\hat{\beta}_j \pm 1{,}96 \cdot \text{ep}(\hat{\beta}_j)$. De fato, quando $n - k - 1 > 50$, o valor de c está próximo demais de 2, de modo que podemos usar uma *regra de bolso* simples para intervalos de confiança de 95%: $\hat{\beta}_j$ mais ou menos duas vezes seu desvio-padrão. Para graus de liberdade pequenos, os percentis exatos devem ser obtidos das tabelas t.

É fácil construir intervalos de confiança para qualquer outro nível de confiança. Por exemplo, um IC de 90% é obtido ao escolher c como o 95º percentil da distribuição t_{n-k-1}. Quando $gl = n - k - 1 = 25$, $c = 1{,}71$, e portanto o IC de 90% é $\hat{\beta}_j \pm 1{,}71 \cdot \text{ep}(\hat{\beta}_j)$, que é necessariamente mais estreito que o IC de 95%. Para um IC de 99%, c é o 99,5º percentil da distribuição t_{25}. Com $gl = 25$, o IC de 99% é aproximadamente $\hat{\beta}_j \pm 2{,}79 \cdot \text{ep}(\hat{\beta}_j)$, que é inevitavelmente mais largo que o IC de 95%.

Muitos programas de regressão modernos poupam-nos de fazer quaisquer cálculos ao informar um IC de 95% juntamente com cada coeficiente e seu erro-padrão. Visto que um intervalo de confiança é construído, é fácil realizar um teste de hipóteses bicaudal. Se a hipótese nula for H_0: $\beta_j = a_j$, então H_0 é rejeitada contra H_1: $\beta_j \neq a_j$ ao nível de significância de (por exemplo) 5% se, e somente se, a_j *não* está no intervalo de confiança de 95%.

EXEMPLO 4.8

(Modelo de Preço Hedônico de Casas)

Um modelo que explica o preço de um bem em termos das características desse bem é chamado *modelo de preço hedônico*. A equação seguinte é um modelo de preço hedônico de preços de casas; as características são área (*arquad*), número de quartos (*qtdorm*) e número de banheiros (*banhos*). Em geral, *preço* aparece na forma logarítmica, assim como algumas das variáveis explicativas. Usando n = 19 observações sobre casas que foram vendidas em Waltham, Massachusetts, em 1990, a equação estimada (com os erros-padrão entre parênteses abaixo das estimativas dos coeficientes) é

$$\log(\hat{preço}) = 7{,}46 + 0{,}634\log(arquad) - 0{,}066\, qtdorm + 0{,}158\, banhos$$
$$(1{,}15)\ (0{,}184)\qquad (0{,}059)\qquad (0{,}075)$$
$$n = 19,\ R^2 = 0{,}806.$$

Como tanto *preço* como *arquad* aparecem na forma logarítmica, a elasticidade preço em relação à área é 0,634, de modo que, mantendo o número de quartos e banheiros fixos, um aumento de 1% na área aumenta o preço previsto da casa em cerca de 0,634%. Podemos construir um intervalo de confiança de 95% para a elasticidade populacional usando o fato de que o modelo estimado tem $n - k - 1 = 19 - 3 - 1 = 15$ graus de liberdade. Da Tabela G.2, achamos o 97,5º percentil de uma distribuição t_{25}: c = 2,131. Assim, o intervalo de confiança para $\beta_{\log(arquad)}$ é 0,634 ± 2,131(0,184), ou (0,242, 1,026). Como zero está excluído desse intervalo de confiança, rejeitamos H_0: $\beta_{\log(arquad)} = 0$ contra a alternativa bilateral ao nível de 5%.

O coeficiente de *qtdorm* é negativo, o que parece contra-intuitivo. Entretanto, é importante lembrar a natureza *ceteris paribus* do coeficiente: ele mede o efeito de mais um quarto, mantendo fixos o tamanho da casa e o número de banheiros. Se duas casas têm o mesmo tamanho, mas uma tem mais quartos, então a casa com mais quartos tem quartos menores; mais quartos menores não é, necessariamente, uma coisa boa. Em qualquer caso, podemos ver que o intervalo de confiança de 95% para β_{qtdorm} é um pouco largo, e ele contém o valor zero: −0,66 ± 2,131(0,059) ou (−0,192, 0,060). Assim, *quartos* não tem um efeito *ceteris paribus* estatisticamente significante sobre o preço das casas.

Dados o tamanho e o número de quartos, prevê-se que um banheiro a mais aumenta o preço da casa em cerca de 15,8%. (Lembre-se de que devemos multiplicar o coeficiente de *banhos* por 100 para obter o efeito em percentagem.) O intervalo de confiança de 95% para β_{banhos} é (−0,002,0,318). Nesse caso, zero está por muito pouco dentro do intervalo de confiança; assim, tecnicamente falando, β_{banhos} não é estatisticamente significante ao nível de 5% contra uma hipótese alternativa bilateral. Como ele está muito próximo de ser significante, provavelmente concluiríamos que o número de banheiros tem um efeito sobre log(*preço*).

Você deve lembrar que um intervalo de confiança é tão bom quanto as hipóteses subjacentes feitas para construí-lo. Se omitirmos fatores importantes que são correlacionados com as variáveis explicativas, então as estimativas dos coeficientes não são confiáveis: MQO é viesado. Se a heteroscedasticidade está presente — por exemplo, no exemplo anterior, se a variância de log(*preço*) depende de qualquer uma das variáveis explicativas —, então o erro-padrão não é válido como uma estimativa de ep $(\hat{\beta}_j)$ (como discutido na Seção 3.4), e o intervalo de confiança calculado ao se usar esses erros-padrão não será, verdadeiramente, um IC de 95%. Também usamos a hipótese de normalidade dos erros para obter esses ICs mas, como veremos no Capítulo 5, isso não é tão importante para aplicações que envolvem centenas de observações.

4.4 TESTES DE HIPÓTESES SOBRE UMA COMBINAÇÃO LINEAR DOS PARÂMETROS

As duas seções anteriores mostraram como usar o teste de hipóteses clássico ou os intervalos de confiança para testar hipóteses sobre um único β_j de cada vez. Nas aplicações, devemos freqüentemente testar hipóteses que envolvem mais de um dos parâmetros da população. Nesta seção, vamos mostrar como testar uma única hipótese envolvendo mais de um dos β_j. A Seção 4.5 mostrará como testar hipóteses múltiplas.

Para ilustrar a abordagem geral, consideraremos um modelo simples para comparar os retornos da educação de cursos superiores profissionalizantes de dois anos (*junior colleges*) e de cursos superiores de quatro anos (*four-year colleges*); por simplicidade, vamos nos referir ao último como "universidades". [Kane e Rouse (1995) fazem uma análise detalhada dos retornos dos *junior colleges* e dos *four-year colleges*.] A população inclui as pessoas com o ensino médio completo que trabalham, e o modelo é

$$\log(salário) = \beta_0 + \beta_1 cp; \beta_2 univ + \beta_3 exper + u, \tag{4.17}$$

em que *cp* é o número de anos freqüentados em um curso superior profissionalizante de dois anos e *univ* é o número de anos freqüentados em um curso superior de quatro anos. Note que qualquer combinação de curso profissionalizante e curso de quatro anos é permitida, incluindo $cp = 0$ e $univ = 0$.

A hipótese de interesse é se um ano no curso profissionalizante é equivalente a um ano na universidade: isso é expresso como

$$H_0: \beta_1 = \beta_2. \tag{4.18}$$

Sob H_0, um ano a mais no curso profissionalizante e um ano a mais na universidade levam ao mesmo aumento percentual *ceteris paribus* em *salário*. Na maioria dos casos, a alternativa de interesse é unilateral: um ano no curso profissionalizante é menos valioso do que um ano na universidade. Isso é expresso como

$$H_1: \beta_1 < \beta_2. \tag{4.19}$$

As hipóteses (4.18) e (4.19) dizem respeito a *dois* parâmetros, β_1 e β_2, uma situação com a qual não tínhamos nos deparado ainda. Não podemos simplesmente usar as estatísticas *t* individuais de $\hat{\beta}_1$ e $\hat{\beta}_2$ para testar H_0. Entretanto, conceitualmente, não há dificuldade em construir uma estatística *t* para testar (4.18). A fim de fazer isso, vamos reescrever a hipótese nula e a alternativa como $H_0: \beta_1 - \beta_2 = 0$ e $H_1: \beta_1 - \beta_2 < 0$, respectivamente. A estatística *t* é baseada em se a diferença estimada $\hat{\beta}_1 - \hat{\beta}_2$ é suficientemente menor que zero para assegurar a rejeição de (4.18) em favor de (4.19). Para considerar o erro de nossos estimadores, padronizamos essa diferença ao dividi-la pelo erro-padrão:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\text{ep}(\hat{\beta}_1 - \hat{\beta}_2)}. \tag{4.20}$$

Uma vez que temos a estatística t de (4.20), o teste segue o procedimento anterior. Escolhemos um nível de significância para o teste e, com base nos gl, obtemos um valor crítico. Como a alternativa é da forma (4.19), a regra de rejeição é da forma $t < -c$, em que c é um valor positivo escolhido de uma distribuição t apropriada. Ou então calculamos a estatística t e, em seguida, o p-valor (veja a Seção 4.2).

A única coisa que faz com que o teste da igualdade de dois parâmetros diferentes seja mais difícil do que testar um único β_j é a obtenção do erro-padrão do denominador de (4.20). Obter o numerador é trivial, uma vez que tenhamos computado a regressão de MQO. Ao usar os dados do arquivo TWOYEAR.RAW, provenientes de Kane e Rouse (1995), estimamos a equação (4.17):

$$\begin{aligned} \log(\widehat{salário}) &= 1{,}472 + 0{,}0667\, cp + 0{,}0769\, univ + 0{,}0049\, exper \\ &\qquad (0{,}021)\quad (0{,}0068)\quad\ (0{,}0023)\quad\ (0{,}0002) \\ &n = 6{,}763,\ R^2 = 0{,}222. \end{aligned} \tag{4.21}$$

De (4.21), fica claro que *cp* e *univ* têm ambos os efeitos – econômico e estatístico – significantes sobre o salário. Isso é, certamente, de interesse, mas estamos mais interessados em testar se a *diferença* estimada dos coeficientes é estatisticamente significante. A diferença é estimada como $\hat{\beta}_1 - \hat{\beta}_2 = -0{,}0102$, de modo que o retorno de um ano em um curso profissionalizante é cerca de um ponto percentual menor que um ano na universidade. Economicamente, isso não é uma diferença trivial. A diferença de $-0{,}0102$ é o numerador da estatística t em (4.20).

Infelizmente, os resultados da regressão em (4.21) *não* contêm informações suficientes para obter o erro-padrão de $\hat{\beta}_1 - \hat{\beta}_2$. Pode ser tentador afirmar que $\text{ep}(\hat{\beta}_1 - \hat{\beta}_2) = \text{ep}(\hat{\beta}_1) - \text{ep}(\hat{\beta}_2)$, mas isso não é verdade. De fato, se invertêssemos os papéis de $\hat{\beta}_1$ e $\hat{\beta}_2$, terminaríamos com um erro-padrão negativo da diferença ao usar a diferença dos erros-padrão. Estes devem *sempre* ser positivos porque eles são estimativas dos desvios-padrão. Embora o erro-padrão da diferença $\hat{\beta}_1 - \hat{\beta}_2$ dependa, certamente, de $\text{ep}(\hat{\beta}_1)$ e $\text{ep}(\hat{\beta}_2)$, ele depende de uma maneira um pouco complicada. Para encontrar $\text{ep}(\hat{\beta}_1 - \hat{\beta}_2)$, primeiro obtemos a variância da diferença. Ao usar os resultados das variâncias do Apêndice B (disponível no site da Thomson), temos

$$\text{Var}(\hat{\beta}_1 - \hat{\beta}_2) = \text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) - 2\,\text{Cov}(\hat{\beta}_1, \hat{\beta}_2). \tag{4.22}$$

Observe, cuidadosamente, como as duas variâncias são *somadas* e, então, a covariância é subtraída duas vezes. O desvio-padrão de $\hat{\beta}_1 - \hat{\beta}_2$ é exatamente a raiz quadrada de (4.22), e, como $[\text{ep}(\hat{\beta}_1)]^2$ é um estimador não-viesado de $\text{Var}(\hat{\beta}_1)$, e similarmente para $[\text{ep}(\hat{\beta}_2)]^2$, temos

$$\text{ep}(\hat{\beta}_1 - \hat{\beta}_2) = \{[\text{ep}(\hat{\beta}_1)]^2 + [\text{ep}(\hat{\beta}_2)]^2 - 2s_{12}\}^{1/2}, \tag{4.23}$$

em que s_{12} é uma estimativa de $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$. Não mostramos uma fórmula para $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$. Alguns programas de regressão têm características que nos permitem obter s_{12}, caso em que se pode calcular o erro-padrão em (4.23) e, em seguida, a estatística t em (4.20). O Apêndice E mostra com usar a álgebra matricial para obter s_{12}.

Vamos sugerir uma outra rota que é muito mais simples de calcular, menos provável de levar a erro e prontamente aplicável a uma variedade de problemas. Em vez de tentar calcular $\text{ep}(\hat{\beta}_1 - \hat{\beta}_2)$ a partir de (4.23), é muito mais fácil estimar um modelo diferente que produz, diretamente, o erro-padrão

de interesse. Defina um novo parâmetro como a diferença entre β_1 e β_2: $\theta_1 = \beta_1 - \beta_2$. Então, queremos testar

$$H_0: \theta_1 = 0 \text{ contra } H_1: \theta_1 < 0. \tag{4.24}$$

A estatística t em (4.20), em termos de $\hat{\theta}_1$, é exatamente $t = \hat{\theta}_1/\text{ep}(\hat{\theta}_1)$. O desafio é encontrar $\text{ep}(\hat{\theta}_1)$.

Podemos obter isso ao reescrever o modelo, de modo que θ_1 aparece diretamente como o coeficiente de uma das variáveis independentes. Como $\theta_1 = \beta_1 - \beta_2$, podemos também escrever $\beta_1 = \theta_1 + \beta_2$. Inserindo em (4.17) e rearranjando, resulta a equação

$$\begin{aligned} \log(\textit{salário}) &= \beta_0 + (\theta_1 + \beta_2)\textit{cp} + \beta_2\textit{univ} + \beta_3\textit{exper} + u \\ &= \beta_0 + \theta_1\textit{cp} + \beta_2(\textit{cp} + \textit{univ}) + \beta_3\textit{exper} + u. \end{aligned} \tag{4.25}$$

A idéia fundamental é que o parâmetro θ_1, cuja hipótese estamos interessados em testar, multiplica agora a variável *cp*. O intercepto ainda é β_0, e *exper* também aparece multiplicado por β_3. Mais importante, há uma nova variável multiplicando β_2, a saber, *pc* + *univ*. Assim, se quisermos estimar diretamente θ_1 e obter o erro-padrão $\hat{\theta}_1$, então devemos construir a nova variável *pc* + *univ* e incluí-la no modelo de regressão no lugar de *univ*. Nesse exemplo, a nova variável tem uma interpretação natural: ela é o *total* de anos de curso superior; assim, defina *totgrad* = *pc* + *univ* e escreva (4.25) como

$$\log(\textit{salário}) = \beta_0 + \theta_1\textit{cp} + \beta_2\textit{totgrad} + \beta_3\textit{exper} + u. \tag{4.26}$$

O parâmetro β_1 desapareceu do modelo, enquanto θ_1 aparece explicitamente. Esse modelo é, de fato, uma maneira diferente de escrever o modelo original. A única razão pela qual definimos esse modelo é que, quando o estimamos, o coeficiente de *cp* ($\hat{\theta}_1$), e, mais importante, $\text{ep}(\hat{\theta}_1)$ é informado juntamente com a estimativa. A estatística t que queremos é a que está relacionada a *cp* (e *não* à *totgrad*), e é informada por qualquer programa de regressão.

Quando fazemos isso com as 6.763 observações utilizadas anteriormente, o resultado é

$$\begin{aligned} \log(\widehat{\textit{salário}}) &= 1{,}472 - 0{,}0102\,\textit{cp} + 0{,}0769\,\textit{totgrad} + 0{,}0049\,\textit{exper} \\ &\quad (0{,}021)\ (0{,}0069) \qquad (0{,}0023) \qquad\quad (0{,}0002) \\ &\quad n = 6{,}763,\ R^2 = 0{,}222. \end{aligned} \tag{4.27}$$

Nessa equação, o único número que não poderíamos obter de (4.21) é o erro-padrão da estimativa −0,0102, que é igual a 0,0069. A estatística t para testar (4,18) é −0,0102/0,0069 = −1,48. Contra a alternativa unilateral (4.19), o p-valor é cerca de 0,070; assim, há alguma, mas não forte, evidência contra (4.18).

O intercepto e a estimativa de inclinação de *exper*, juntamente com os erros-padrão, são os mesmos de (4.21). Esse fato *deve* ser exato, e ele fornece uma maneira de checar se a equação transformada foi apropriadamente estimada. O coeficiente da nova variável, *totgrad*, é a mesma do coeficiente de *univ* em (4.21), e o erro-padrão também é o mesmo. Sabemos que isso deve acontecer ao comparar (4.17) e (4.25).

É bastante simples calcular um intervalo de confiança de 95% para $\theta_1 = \beta_1 - \beta_2$. Usando a aproximação normal padronizada, o IC é obtido da maneira usual: $\hat{\theta}_1 \pm 1{,}96\ \text{ep}(\hat{\theta}_1)$, que, nesse caso, leva a $-0{,}0102 \pm 0{,}0135$.

A estratégia de reescrever o modelo, de modo que ele contenha o parâmetro de interesse, funciona em todos os casos e é fácil de implementar. (Veja os problemas 4.12 e 4.14 para outros exemplos.)

4.5 TESTES DE RESTRIÇÕES LINEARES MÚLTIPLAS: O TESTE *F*

A estatística *t* associada com qualquer coeficiente de MQO pode ser usada para testar se o parâmetro desconhecido correspondente na população é igual a qualquer constante dada (freqüentemente, mas não sempre, zero). Acabamos de mostrar como testar hipóteses sobre uma única combinação linear dos β_j ao rearranjar a equação e computar uma regressão usando variáveis transformadas. No entanto, até agora, somente cobrimos hipóteses que envolvem uma *única* restrição. Freqüentemente, desejamos testar hipóteses *múltiplas* sobre os parâmetros subjacentes $\beta_0, \beta_1, \ldots, \beta_k$. Vamos começar com o caso de testar se um conjunto de variáveis independentes não tem efeito parcial sobre uma variável dependente.

Teste de Restrições de Exclusão

Já sabemos como testar se uma variável particular não tem efeito sobre a variável dependente, usando a estatística *t*. Agora, queremos testar se um *grupo* de variáveis não tem efeito sobre a variável dependente. Mais precisamente, a hipótese nula é que um conjunto de variáveis não tem efeito sobre *y*, já que outro conjunto de variáveis foi controlado.

Como uma ilustração do porquê de testar a significância de um grupo é útil, vamos considerar o seguinte modelo que explica os salários dos jogadores da principal liga de beisebol dos Estados Unidos:

$$\log(salário) = \beta_0 + \beta_1 anos + \beta_2 jogosano + \beta_3 rebmed + \beta_4 hrunano + \beta_5 rebrunano + u, \tag{4.28}$$

em que *salário* é o salário total do jogador em 1993, *anos* corresponde aos anos do jogador na liga, *jogosano* é a média de partidas jogadas por ano, *rebmed* é a média de rebatidas na carreira do jogador, *hrunano* corresponde a rebatidas que redundaram em pontos (volta completa por todas as bases) por ano, e *rebrunano* corresponde a rebatidas que redundaram em corrida até a próxima base por ano. Suponha que queiramos testar a hipótese nula de que, uma vez que anos na liga e jogos por ano foram controlados, as estatísticas que medem o desempenho – *rebmed*, *hrunano* e *rebrunano* – não têm efeito sobre o salário. Essencialmente, a hipótese nula expressa que a produtividade, medida pelas estatísticas do beisebol, não tem efeito sobre o salário.

Em termos dos parâmetros do modelo, a hipótese nula é formulada como

$$H_0: \beta_3 = 0, \beta_4 = 0, \beta_5 = 0. \tag{4.29}$$

A hipótese nula (4.29) constitui três **restrições de exclusão**: se (4.29) é verdadeira, então *rebmed*, *hrunano* e *rebrunano* não têm efeito sobre log(*salário*) após *anos* e *jogosano* terem sido controlados e, portanto, deveriam ser excluídos do modelo. Esse é um exemplo de conjunto de **restrições múltiplas**

porque estamos colocando mais de uma restrição sobre os parâmetros de (4.28); posteriormente, veremos mais exemplos gerais de restrições múltiplas. Um teste de restrições múltiplas é chamado **teste de hipóteses múltiplas** ou o **teste de hipóteses conjuntas.**

Qual seria a alternativa a (4.29)? Se o que temos em mente é que "estatísticas de desempenho importam, mesmo após controlar as variáveis anos na liga e jogos por ano", então a hipótese alternativa é simplesmente

$$H_1: H_0 \text{ não é verdadeira.} \qquad \textbf{(4.30)}$$

A alternativa (4.30) se mantém quando pelo menos um dos β_3, β_4 ou β_5 for diferente de zero. (Qualquer um deles ou todos poderiam ser diferentes de zero.) O teste que estudamos aqui é construído para detectar qualquer violação de H_0. Ele também é válido quando a hipótese alternativa é algo como H_1: $\beta_3 > 0$, ou $\beta_4 > 0$, ou $\beta_5 > 0$, mas ele não será o melhor teste possível sob essas alternativas. Não temos o espaço ou a formação estatística necessários para cobrir testes mais poderosos sob hipóteses alternativas unilaterais múltiplas.

Como devemos proceder para testar (4.29) contra (4.30)? É tentador testar (4.29) usando as estatísticas t das variáveis *rebmed*, *hrunano* e *rebrunano* para determinar se cada variável é *individualmente* significante. Essa opção não é apropriada. Uma estatística t particular testa uma hipótese que não coloca restrições sobre os outros parâmetros. Além disso, teríamos três resultados para resolver o problema – um para cada estatística t. Qual deles constituiria a rejeição de (4.29) ao nível de, por exemplo, 5%? Dever-se-ia exigir que todas as três estatísticas t são significantes ao nível de 5% ou somente uma das três? Essas são questões difíceis, e felizmente não temos de respondê-las. Além do mais, usar estatísticas t separadas para testar uma hipótese múltipla como (4.29) pode ser muito enganoso. Precisamos de uma maneira para testar as restrições de exclusão *conjuntamente*.

Para ilustrar essas questões, estimamos a equação (4.28) usando os dados do arquivo MLB1.RAW. Obtemos

$$\begin{aligned} \log(\widehat{salary}) &= 11{,}10 + 0{,}0689\,anos + 0{,}0126\,jogosano \\ &\qquad\quad (0{,}29)\quad (0{,}0121)\qquad (0{,}0026) \\ &+ 0{,}00098\,rebmed + 0{,}0144\,hrunano + 0{,}0108\,rebrunano \\ &\qquad\quad (0{,}00110)\qquad (0{,}0161)\qquad (0{,}0072) \\ n &= 353,\ \text{SQR} = 183{,}186,\ R^2 = 0{,}6278, \end{aligned} \qquad \textbf{(4.31)}$$

em que SQR é a soma dos resíduos quadrados. (vamos usá-lo mais tarde.) A fim de facilitar futuras comparações, deixamos vários números após a vírgula em SQR e no R-quadrado. A equação (4.31) revela que, enquanto *anos* e *jogosano* são estatisticamente significantes, nenhuma das variáveis *rebmed*, *hrunano* e *rebrunano* tem uma estatística t estatisticamente significante contra uma alternativa bilateral ao nível de significância de 5%. (A estatística t de *rebrunano* está muito próxima de ser significante; seu p-valor bilateral é 0,134.) Assim, baseados nas três estatísticas t, parece que não podemos rejeitar H_0.

Essa conclusão revela-se errada. A fim de ver isso, devemos derivar um teste de restrição múltipla cuja distribuição seja conhecida e tabelada. A soma dos resíduos quadrados aparece, agora, para dar uma base muito conveniente para testar hipóteses múltiplas. Também mostraremos como o R-quadrado pode ser usado no caso especial de testar restrições de exclusão.

Conhecer a soma dos resíduos quadrados em (4.31) não nos diz nada sobre a decisão quanto à hipótese nula em (4.29). O que nos dirá algo é: saber de quanto aumenta SQR quando retiramos as variáveis *rebmed*, *hrunano* e *rebrunano* do modelo. Lembre-se de que, como as estimativas de MQO são escolhidas para minimizar a soma dos resíduos quadrados, o SQR *sempre* aumenta quando variáveis são retiradas do modelo; esse é um fato algébrico. A questão é saber se esse aumento é suficientemente grande, *relativamente* ao SQR do modelo com todas as variáveis, para garantir a rejeição da hipótese nula.

O modelo sem as três variáveis em questão é simplesmente

$$\log(salário) = \beta_0 + \beta_1 anos + \beta_2 jogosano + u. \qquad \textbf{(4.32)}$$

No contexto do teste de hipóteses, a equação (4.32) é o **modelo restrito** para testar (4.29); o modelo (4.28) é chamado **modelo irrestrito**. O modelo restrito sempre tem menos parâmetros que o modelo irrestrito.

Quando estimamos o modelo restrito usando os dados do arquivo MLB1.RAW, obtemos

$$\begin{aligned} \log(\hat{salário}) &= 11{,}22 + 0{,}0713\, anos + 0{,}0202\, jogosano \\ &\qquad (0{,}11)\quad (0{,}0125)\qquad\quad (0{,}0013) \\ n &= 353,\ \text{SQR} = 198{,}311,\ R^2 = 0{,}5971. \end{aligned} \qquad \textbf{(4.33)}$$

Como imaginamos, o SQR de (4.33) é maior que o SQR de (4.31), e o R-quadrado do modelo restrito é menor que o R-quadrado do modelo irrestrito. O que precisamos decidir é se, ao passarmos do modelo irrestrito para o modelo restrito, o aumento em SQR (183,186 para 198,311) é suficientemente grande para garantir a rejeição de (4.29). Como em todo teste, a resposta depende do nível de significância do teste. No entanto, não podemos realizar o teste a um determinado nível de significância até que tenhamos uma estatística cuja distribuição seja conhecida, e possa ser tabelada, sob H_0. Assim, precisamos de uma maneira de combinar as informações dos dois SQRs para obter uma estatística de teste com uma distribuição conhecida sob H_0.

Podemos derivar o teste para o caso geral, visto que isso não é tão difícil. Escreva o modelo *irrestrito* com k variáveis independentes como

$$y = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k + u; \qquad \textbf{(4.34)}$$

o número de parâmetros no modelo irrestrito é $k + 1$. (Lembre-se de adicionar um por causa do intercepto.) Suponha que temos q restrições de exclusão para testar: isto é, a hipótese nula afirma que q variáveis em (4.34) têm coeficientes zero. Por simplicidade notacional, assuma que sejam as q últimas variáveis da lista de variáveis independentes: $x_{k-q+1}, \ldots, x_k$. (A ordem das variáveis, evidentemente, é arbitrária e não importa.) A hipótese nula é formulada como

$$H_0: \beta_{k-q+1} = 0, \ldots, \beta_k = 0, \qquad \textbf{(4.35)}$$

que coloca q restrições de exclusão sobre o modelo (4.34). A hipótese alternativa a (4.35) é simplesmente que H_0 é falsa; isso significa que pelo menos um dos parâmetros listados em (4.35) é diferente de zero. Quando impomos as restrições sob H_0, ficamos com o modelo restrito:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_{k-q} x_{k-q} + u. \qquad \textbf{(4.36)}$$

Nesta seção, vamos assumir que ambos os modelos irrestrito e restrito contêm um intercepto, já que esse é o caso mais amplamente encontrado na prática.

Agora, vamos à estatística de teste propriamente dita. Anteriormente, sugerimos que olhar para o aumento relativo em SQR quando nos movemos do modelo irrestrito para o restrito deveria ser informativo para testar a hipótese (4.35). A **estatística *F*** (ou razão *F*) é definida como

$$F \equiv \frac{(SQR_r - SQR_{ir})/q}{SQR_{ir}/(n-k-1)}, \qquad \textbf{(4.37)}$$

em que SQR_r é a soma dos resíduos quadrados do modelo restrito, e SQR_{ir} é a soma dos resíduos quadrados do modelo irrestrito.

Você deveria observar imediatamente que, como SQR_r não pode ser maior que SQR_{ir}, a estatística F é *sempre* não-negativa (e quase sempre estritamente positiva). Assim, se você calcular uma estatística F negativa, algo está errado; em geral, a ordem dos SQRs no numerador de F é equivocadamente invertida. Também, o SQR no denominador de F é o SQR do modelo irrestrito. A maneira mais fácil de lembrar onde os SQRs aparecem é pensar em F medindo o aumento relativo em SQR quando nos movemos do modelo irrestrito para o restrito.

Considere a possibilidade de relacionar o desempenho individual em um teste padronizado, *nota*, a uma variedade de outras variáveis. Fatores relativos à escola incluem o tamanho médio da classe, os gastos por estudante, o salário médio dos professores e o total de matrículas escolares. Outras variáveis específicas aos estudantes são a renda familiar, a educação da mãe, a educação do pai e o número de irmãos. O modelo é

$$\begin{aligned} nota = {} & \beta_0 + \beta_1 tclasse + \beta_2 gasto + \beta_3 totalsalp + \\ & \beta_4 matricl + \beta_5 rendfam + \beta_6 educm + \\ & \beta_7 educp + \beta_8 irmãos + u. \end{aligned}$$

Formule a hipótese nula de que as variáveis específicas aos estudantes não têm efeito sobre o desempenho no teste padronizado, uma vez que os fatores relativos à escola sejam controlados. Quais são os valores de k e q nesse exemplo? Escreva a versão restrita do modelo.

A diferença nos SQRs no numerador de F é dividida por q, o qual é o número de restrições impostas ao nos movermos do modelo irrestrito para o restrito (q variáveis independentes foram retiradas). Portanto, podemos escrever

$$q = \textbf{graus de liberdade do numerador} = gl_r - gl_{ir}, \tag{4.38}$$

o que também mostra que q é a diferença nos graus de liberdade entre os modelos restrito e irrestrito. (Lembre-se de que gl = número de observações-número de parâmetros estimados.) Visto que o modelo restrito tem menos parâmetros — e cada modelo é estimado usando as mesmas n observações —, gl_r é sempre maior que gl_{ir}.

O SQR no denominador de F é dividido pelos graus de liberdade do modelo irrestrito:

$$n - k - 1 = \textbf{graus de liberdade do denominador} = gl_{ir}. \tag{4.39}$$

De fato, o denominador de F é exatamente o estimador não-viesado de $\sigma^2 = \text{Var}(u)$ do modelo irrestrito.

Em uma aplicação particular, calcular a estatística F é mais fácil do que ler penosamente a notação um pouco incômoda usada para descrever o caso geral. Em primeiro lugar, obtemos os graus de liberdade do modelo irrestrito, gl_{ir}. Então, contamos quantas variáveis estão excluídas no modelo restrito; esse é o valor de q. Os SQRs são informados em toda regressão de MQO e, portanto, é simples compor a estatística F.

Na regressão do salário da principal liga de beisebol, $n = 353$, e o modelo completo (4.28) contém seis parâmetros. Assim, $n - k - 1 = gl_{ir} = 353 - 6 = 347$. O modelo restrito (4.32) contém menos três variáveis independentes que (4.28), e portanto, $q = 3$. Assim, temos todos os ingredientes para calcular a estatística F; vamos adiar o cálculo até que saibamos o que fazer com ele.

A fim de usar a estatística F, devemos conhecer sua distribuição amostral sob a hipótese nula para escolher os valores críticos e as regras de rejeição. Pode ser mostrado que, sob H_0 (e assumindo que as hipóteses do MCL se mantêm), F é distribuído como uma variável aleatória F com $(q, n - k - 1)$ graus de liberdade. Escrevemos isso como

$$F \sim F_{q,\, n-k-1}.$$

A distribuição de $F_{q,\, n-k-1}$ está tabelada e disponível em tabelas estatísticas (veja a Tabela G.3) e, ainda mais importante, em programas estatísticos.

Não vamos derivar a distribuição F porque a matemática é muito complicada. Basicamente, pode ser mostrado que a equação (4.37) é, de fato, a razão de duas variáveis aleatórias qui-quadradas independentes, divididas por seus respectivos graus de liberdade. A variável aleatória qui-quadrada do numerador tem q graus de liberdade, e a qui-quadrada do denominador tem $n - k - 1$ graus de liberdade. Essa é a definição de uma variável aleatória com distribuição F (veja o Apêndice B, disponível no site da Thomson).

Da definição de F, é bastante claro que rejeitaremos H_0 em favor de H_1 quando F for suficientemente "grande". A grandeza depende de nosso nível de significância escolhido. Suponha que decidimos por um teste ao nível de 5%. Seja c o 95º percentil da distribuição $F_{q,\, n-k-1}$. O valor crítico depende de q (os gl do numerador) e $n - k - 1$ (os gl do denominador). É importante guardar corretamente os graus de liberdade do numerador e do denominador.

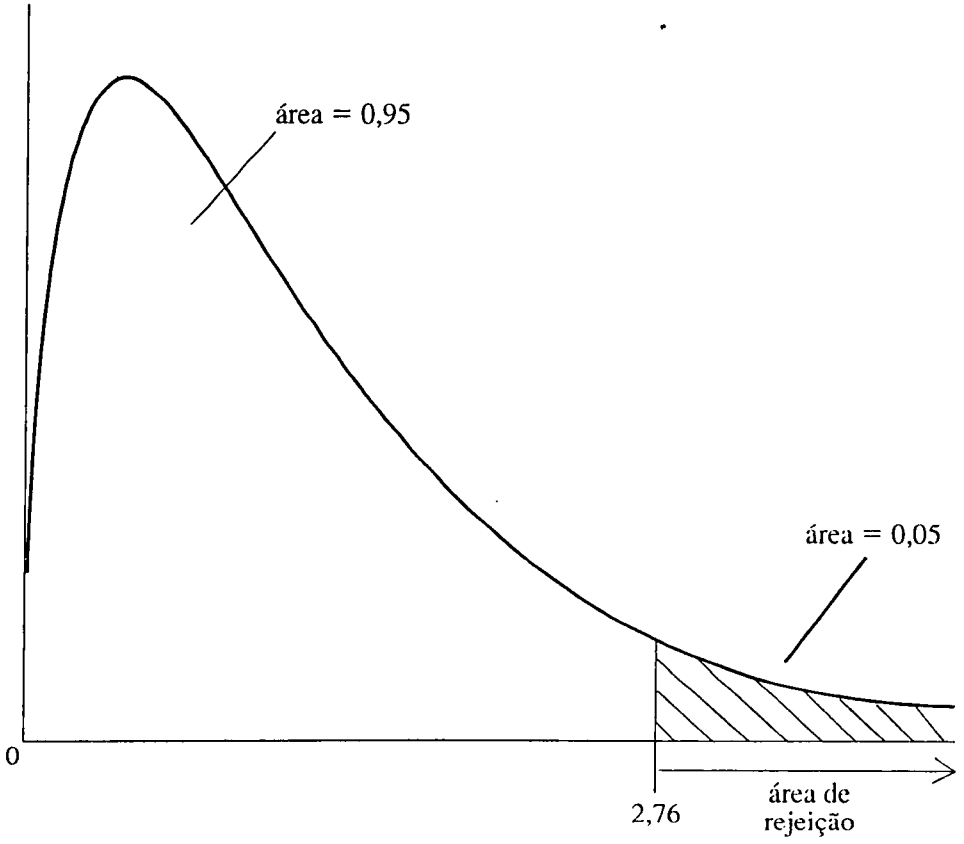
Os valores críticos de 10%, 5% e 1% da distribuição F são dados na Tabela 1.3. A regra de rejeição é simples. Uma vez obtido c, rejeitamos H_0 em favor de H_1, ao nível de significância escolhido se

$$F > c. \tag{4.40}$$

Com um nível de significância de 5%, $q = 3$ e $n - k - 1 = 60$, o valor crítico é $c = 2{,}76$. Rejeitaríamos H_0 ao nível de significância de 5% se o valor calculado da estatística F excedesse 2,76. O valor crítico a 5% e a região de rejeição são apresentados na Figura 4.7. Para os mesmos graus de liberdade, o valor crítico a 1% é 4,13.

Figura 4.7

O valor crítico de 5% e a região de rejeição em uma distribuição $F_{3,60}$.



Na maioria das aplicações, os graus de liberdade do numerador (q) serão notadamente menores que os graus de liberdade do denominador ($n - k - 1$). As aplicações em que $n - k - 1$ menor têm menos probabilidade de serem bem-sucedidas porque os parâmetros do modelo irrestrito provavelmente não serão estimados com precisão. Quando os gl do denominador alcançam cerca de 120, a distribuição F não é mais sensível a eles. (Isso é totalmente semelhante à distribuição t aproximada pela distribuição normal padronizada quando os gl tornam-se grandes.) Assim, há uma entrada na tabela para o denominador $gl = \infty$, e isso é o que usamos com amostras grandes (visto que $n - k - 1$ é, então, grande). Uma formulação semelhante é válida para os gl do numerador grandes, mas isso raramente ocorre nas aplicações.

Se H_0 é rejeitada, dizemos que $x_{k-q+1}, \ldots, x_k$ são **estatisticamente significantes conjuntamente** (ou apenas *conjuntamente significantes*) ao nível de significância apropriado. Esse teste sozinho não nos permite dizer quais das variáveis têm um efeito parcial sobre y; todas elas podem afetar y ou talvez somente uma afeta. Se a hipótese nula não for rejeitada, as variáveis são **conjuntamente não significantes**, o que, em geral, justifica retirá-las do modelo.

No exemplo da principal liga de beisebol com três graus de liberdade do numerador e 347 graus de liberdade do denominador, o valor crítico a 5% é 2,60, e o valor crítico a 1% é 3,78. Rejeitamos H_0 ao nível de 1% se F está acima de 3,78; rejeitamos H_0 ao nível de 5% se F está acima de 2,60.

Estamos agora em posição para testar a hipótese com a qual iniciamos esta seção: após controlar *anos* e *jogosano*, as variáveis *rebmed*, *hrunano* e *rebrunano* não têm efeito sobre os salários dos jogadores. Na prática, é mais fácil, em primeiro lugar, calcular $(SQR_r - SQR_{ir})/SQR_{ir}$ e então multiplicar o resultado por $(n - k - 1)/q$; a razão pela qual a fórmula é expressa como em (4.37) é que ela torna mais fácil guardar corretamente os graus de liberdade do numerador e do denominador. Usando os SQRs em (4.31) e (4.33), temos

$$F = \frac{(198,311 - 183,186)}{183,186} \cdot \frac{347}{3} \approx 9,55.$$

Esse número está bem acima do valor crítico de 1% da distribuição F com 3 e 347 graus de liberdade e, portanto, rejeitamos completamente a hipótese de que *rebmed*, *hrunano* e *rebrunano* não têm efeito sobre o salário.

O resultado do teste conjunto pode parecer surpreendente à luz das estatísticas t não significantes das três variáveis. O que está acontece é que as variáveis *hrunano* e *rebrunano* são altamente correlacionadas, e essa multicolinearidade torna difícil descobrir o efeito parcial de cada variável; isso é refletido nas estatísticas t individuais. A estatística F testa se essas variáveis (incluindo *rebmed*) são conjuntamente significantes, e a multicolinearidade entre *hrunano* e *rebrunano* é muito menos relevante para testar essa hipótese. No Problema 4.16, pediremos que você estime novamente o modelo retirando *rebrunano*, caso em que *hrunano* torna-se muito significante. O mesmo é verdadeiro para *rebrunano* quando *hrunano* é retirado do modelo.

A estatística F é freqüentemente útil para testar a exclusão de um grupo de variáveis quando as variáveis do grupo são altamente correlacionadas. Por exemplo, suponha que queiramos testar se o desempenho da empresa afeta os salários dos seus diretores executivos. Há muitas maneiras de medir o desempenho das empresas, e não é claro dizer, antecipadamente, qual medida é a mais importante. Como as medidas de desempenho das empresas são, provavelmente, altamente correlacionadas, esperar encontrar medidas individualmente significantes pode ser pedir demais, devido à multicolinearidade. No entanto, um teste F pode ser usado para determinar se, como um grupo, as variáveis de desempenho das empresas afetam o salário.

Relação entre as Estatísticas *F* e *t*

Vimos nesta seção como a estatística F pode ser usada para testar se um grupo de variáveis deve ser incluído em um modelo. O que aconteceria se aplicássemos a estatística F ao caso de testar a significância de uma *única* variável independente? Esse caso certamente não é excluído pelo desenvolvimento anterior. Por exemplo, podemos descrever a hipótese nula como H_0: $\beta_k = 0$ e $q = 1$ (para testar a única restrição de exclusão, de que x_k pode ser excluído do modelo). Da Seção 4.2, sabemos que a estatística t de β_k pode ser usada para testar essa hipótese. A questão, então, é: temos duas maneiras separadas de testar hipóteses sobre um único coeficiente? A resposta é não. É possível mostrar que a estatística F para testar a exclusão de uma única variável é igual ao *quadrado* da estatística t correspondente. Como t^2_{n-k-1} tem uma distribuição $F_{1,\,n-k-1}$, as duas abordagens levam exatamente ao mesmo resultado, desde que a hipótese alternativa seja bilateral. A estatística t é mais flexível para testar uma única hipótese porque ela pode ser usada para testar alternativas unilaterais. Visto que as estatísticas t também são mais fáceis de serem obtidas do que as estatísticas F, não há razão para usar uma estatística F para testar hipóteses sobre um único parâmetro.

Na regressão dos jogadores da principal liga de beisebol, vimos que duas (ou mais) variáveis que têm, cada uma, estatísticas t não significantes podem ser conjuntamente muito significantes. Também é possível que, em um grupo de muitas variáveis explicativas, uma variável tenha uma estatística t significante, mas o grupo de variáveis é conjuntamente não significante aos níveis de significância usuais. O que devemos fazer com esse tipo de resultado? Em termos concretos, suponha que, em um modelo com muitas variáveis explicativas, não possamos rejeitar a hipótese nula de que β_1, β_2, β_3, β_4 e β_5 são todos iguais a zero ao nível de 5%, ainda que a estatística t de $\hat{\beta}_1$ seja significante ao nível de 5%. Logicamente, não podemos ter $\beta_1 \neq 0$ e também ter β_1, β_2, β_3, β_4 e β_5 todos iguais a zero! Contudo, quando se trata de fazer um teste, é possível que agrupemos um punhado de variáveis não significantes juntamente com uma variável significante e concluamos que o conjunto inteiro de variáveis é conjuntamente não significante. (Tais possíveis conflitos entre um teste t e um teste F conjunto dão outro exemplo da razão de não devermos "aceitar" hipóteses nulas; podemos somente não rejeitá-las.) Espera-se que a estatística F revele se qualquer combinação de um conjunto de coeficientes é diferente de zero, mas ele nunca é o melhor teste para determinar se um único coeficiente é diferente de zero. O teste t é o mais apropriado para testar uma única hipótese. (Tecnicamente, uma estatística F para restrições conjuntas que incluam $\beta_1 = 0$ tem menos poder de detectar $\beta_1 \neq 0$ do que a estatística t usual. Para uma discussão do poder de um teste, veja a Seção C.6 do Apêndice C, disponível no site da Thomson.)

Infelizmente, o fato de podermos às vezes ocultar uma variável estatisticamente significante entre algumas variáveis não significantes pode levar a equívocos se os resultados da regressão não forem cuidadosamente descritos. Por exemplo, suponha que, num estudo dos determinantes das taxas de aprovação de empréstimos de uma cidade, x_1 é a fração de famílias negras na cidade. Suponha que as variáveis x_2, x_3, x_4 e x_5 sejam as frações de famílias chefiadas por diferentes grupos de idade. Ao explicar as taxas de empréstimos, incluiríamos medidas de renda, riqueza, avaliação de crédito, e assim por diante. Suponha que a idade do chefe de família não tenha efeito sobre as taxas de aprovação de empréstimos, uma vez que as outras variáveis sejam controladas. Mesmo se a raça tiver um efeito marginalmente significante, é possível que as variáveis raça e idade sejam conjuntamente não significantes. Alguém que queira concluir que raça não é um fator importante poderia simplesmente escrever algo como "As variáveis raça e idade foram acrescentadas à equação, mas elas foram conjuntamente não significantes ao nível de 5%". Felizmente, a revisão atenta impede esses tipos de conclusões enganosas, mas você deve estar consciente de que elas podem ocorrer.

Freqüentemente, quando uma variável é estatisticamente muito significante e ela é testada conjuntamente com outro conjunto de variáveis, o conjunto será conjuntamente significante. Em tais casos, não há mais inconsistência lógica em rejeitar ambas as hipóteses nulas.

A Forma *R*-quadrado da Estatística *F*

Para testar restrições de exclusão é freqüentemente mais conveniente ter uma forma da estatística F que possa ser calculada usando os R-quadrados dos modelos restrito e irrestrito.

Uma razão para isso é que o R-quadrado está sempre entre zero e um, enquanto os SQRs podem ser muito grandes, dependendo da unidade de y, o que faz dos cálculos baseados nos SQRs algo entediante. Usando o fato de que $SQR_r = SQT(1 - R_r^2)$ e $SQR_{ir} = SQT(1 - R_{ir}^2)$, podemos substituir esses termos em (4.37) para obter

$$F \equiv \frac{(R_{ir}^2 - R_r^2)/q}{(1 - R_{ir}^2)/(n - k - 1)} \qquad \textbf{(4.41)}$$

(observe que os termos SQT são cancelados). Isso se chama a **forma R-quadrado da estatística F.** [Neste ponto, você deve ser advertido de que embora a equação (4.41) seja muito conveniente para testar restrições de exclusão, ela não pode ser aplicada para testar todas as restrições lineares. Como veremos ao discutir como testar restrições lineares gerais, a forma soma dos resíduos quadrados da estatística F é, às vezes, necessária.]

Como o R-quadrado é um resultado informado em quase todas as regressões (embora o SQR não seja), é fácil usar os R-quadrados dos modelos irrestrito e restrito para testar a exclusão de algumas variáveis. Atenção particular deve ser colocada à ordem dos R-quadrados do numerador: o R-quadrado *irrestrito* vem primeiro [compare com os SQRs em (4.37)]. Como $R_{ir}^2 > R_r^2$, isso mostra novamente que F sempre será positivo.

Ao usar a forma R-quadrado para testar a exclusão de um conjunto de variáveis, é importante *não* elevar ao quadrado o R-quadrado antes de colocá-lo na fórmula (4.41), pois isso já foi feito. Todas as regressões informam o R^2, e esses números são colocados diretamente em (4.41). No exemplo do salário dos jogadores de beisebol, podemos usar (4.41) para obter a estatística F:

$$F = \frac{(0{,}6278 - 0{,}5971)}{(1 - 0{,}6278)} \cdot \frac{347}{3} \approx 9{,}54,$$

que está muito próxima da que obtivemos anteriormente. (A diferença se deve a erro de arredondamento.)

EXEMPLO 4.9

(Educação dos Pais em uma Equação do Peso de Nascimentos)

Como outro exemplo de cálculo de uma estatística F, considere o seguinte modelo para explicar o peso de recém-nascidos em termos de vários fatores:

$$pesonas = \beta_0 + \beta_1 cigs + \beta_2 ordnas + \beta_3 rendfam + \beta_4 educm + \beta_5 educp + u, \tag{4.42}$$

em que *pesonas* é o peso de nascimento, em libras, *cigs* é o número médio de cigarros que a mãe fumou por dia durante a gravidez, *ordnas* é a ordem de nascimento dessa criança, *rendfam* é a renda familiar anual, *educm* corresponde aos anos de escolaridade formal da mãe e *educp* corresponde aos anos de escolaridade formal do pai. Vamos testar a hipótese nula de que, após controlar *cigs*, *ordnas* e *rendfam*, a educação dos pais não tem efeito sobre o peso de nascimento. Isso é expresso como H_0: $\beta_4 = 0, \beta_5 = 0$, e portanto há $q = 2$ restrições de exclusão para serem testadas. Há $k + 1 = 6$ parâmetros no modelo irrestrito (4.42), de modo que os *gl* do modelo irrestrito são $n - 6$, em que n é o tamanho da amostra.

Vamos testar essa hipótese usando os dados em BWGHT.RAW. Esse conjunto de dados contém informações de 1.388 nascimentos, mas devemos ser cuidadosos ao contar as observações usadas no teste da hipótese nula. Ocorre que em pelo menos uma das variáveis *educm* e *edcup* estão faltando informações de 197 nascimentos na amostra; essas observações não podem ser incluídas ao estimar o modelo irrestrito. Assim, temos realmente $n = 1.191$ observações e, portanto, há $1.191 - 6 = 1.185$ *gl* no modelo irrestrito. Devemos estar seguros de usar essas mesmas 1.191 observações quando estimarmos o modelo restrito (e não o total das 1.388 observações que estão disponíveis). Em geral, ao estimar o modelo restrito para calcular um teste F, devemos usar as mesmas observações para estimar o modelo irrestrito; de outro modo, o teste não é válido. Quando não faltarem dados, isso deixa de ser um problema.

EXEMPLO 4.9 (continuação)

Os *gl* do numerador são iguais a 2, e os *gl* do denominador, a 1.185; da Tabela G.3, o valor crítico a 5% é $c = 3,0$. Por brevidade, em vez de informar os resultados completos, vamos apresentar somente os *R*-quadrados. O *R*-quadrado do modelo completo é $R^2_{ir} = 0,0387$. Quando *educm* e *educp* são retirados da regressão, o *R*-quadrado cai para $R^2_r = 0,0364$. Assim, a estatística *F* é $F = [(0,0387 - 0,0364)/(1 - 0,0387)](1.185/2) = 1,42$; como esse valor está bem abaixo do valor crítico de 5%, não é possível rejeitar H_0. Em outras palavras, *educm* e *educp* são conjuntamente não significantes na equação do peso de nascimento.

Cálculo dos *p*-Valores para Testes *F*

Para apresentar os resultados dos testes *F*, os *p*-valores são especialmente úteis. Como a distribuição *F* depende dos *gl* do numerador e do denominador, é difícil obter uma impressão de quanto é forte ou fraca a evidência contra a hipótese nula simplesmente olhando para o valor da estatística *F* e um ou dois valores críticos.

No contexto do teste *F*, o *p*-valor é definido como

$$p\text{-valor} = P(\mathcal{F} > F), \tag{4.43}$$

em que, para enfatizar, $\mathcal{F}$ representa uma variável aleatória *F* com $(q, n - k - 1)$ graus de liberdade, e *F* é o valor real da estatística de teste. O *p*-valor ainda tem a mesma interpretação que ele tinha para a estatística *t*: ele é a probabilidade de observarmos um valor de *F* pelo menos tão grande quanto aquele que encontramos, *dado* que a hipótese nula é verdadeira. Um *p*-valor pequeno é evidência contra H_0. Por exemplo, o *p*-valor = 0,016 significa que a probabilidade de observarmos um valor de *F* tão grande quanto aquele para o qual a hipótese nula é verdadeira é somente 1,6%; em geral, rejeitamos H_0 em tais casos. Se o *p*-valor = 0,314, então a probabilidade de observarmos um valor da estatística *F* tão grande quanto aquele sob a hipótese nula é 31,4%. A maioria acharia esse valor uma evidência bastante fraca contra H_0.

Os dados do arquivo ATTEND.RAW foram usados para estimar as duas equações

$$\underset{}{taxafreq} = \underset{(2,87)}{47,13} + \underset{(1,09)}{13,37}\, nmgradp$$
$$n = 680,\ R^2 = 0,183,$$

e

$$taxafreq = \underset{(3,88)}{75,70} + \underset{(1,08)}{17,26}\, nmgradp - \underset{(?)}{1,72}\, tac$$
$$n = 680,\ R^2 = 0,291,$$

em que, como sempre, os erros-padrão estão entre parênteses; o erro-padrão de tac está faltando na segunda equação. Qual é a estatística *t* do coeficiente de tac? (*Sugestão*: Primeiro calcule a estatística *F* da significância de tac.)

Assim como com o teste t, uma vez calculado o p-valor, o teste F pode ser realizado para qualquer nível de significância. Por exemplo, se o p-valor = 0,024, rejeitamos H_0 ao nível de significância de 5%, mas não ao nível de 1%.

O p-valor do teste F no Exemplo 4.9 é 0,238, e portanto a hipótese nula de β_{educm} e β_{educp} serem ambos zero não é rejeitada mesmo ao nível de significância de 20%.

Muitos programas econométricos têm recursos embutidos para testar restrições múltiplas de exclusão. Neles, o cálculo computacional tem várias vantagens sobre o cálculo manual das estatísticas: provavelmente cometeremos menos erros, os p-valores são calculados automaticamente e o problema de falta de dados, como no Exemplo 4.9, é tratado sem qualquer trabalho adicional de nossa parte.

A Estatística *F* para a Significância Geral de uma Regressão

Um conjunto especial de restrições de exclusão é rotineiramente testado por muitos programas de regressão. Essas restrições têm a mesma interpretação, independentemente do modelo. No modelo com k variáveis independentes, podemos escrever a hipótese nula como

$$H_0: x_1, x_2, ..., x_k \text{ não ajudam a explicar } y.$$

Essa hipótese nula é, de certa maneira, muito pessimista. Ela afirma que *nenhuma* das variáveis explicativas tem um efeito sobre y. Expressa em termos dos parâmetros, a hipótese nula é que todos os parâmetros de inclinação são zero:

$$H_0: \beta_1 = \beta_2 = ... = \beta_k = 0, \tag{4.44}$$

e a hipótese alternativa é que pelo menos um dos β_j seja diferente de zero. Outra maneira útil de formular a hipótese nula é que H_0: $E(y|x_1, x_2, ..., x_k) = E(y)$, de modo que conhecer os valores de x_1, x_2, ..., x_k não afeta o valor esperado de y.

Há k restrições em (4.44), e quando as impomos, obtemos o modelo restrito

$$y = \beta_0 + u; \tag{4.45}$$

todas as variáveis independentes foram retiradas da equação. Agora, o R-quadrado da estimação de (4.45) é zero; nada da variação em y está sendo explicado porque não há variáveis explicativas. Portanto, a estatística F para testar (4.44) pode ser escrita como

$$\frac{R^2/k}{(1 - R^2)/(n - k - 1)}, \tag{4.46}$$

em que R^2 é exatamente o R-quadrado usual da regressão de y sobre $x_1, x_2, ..., x_k$.

A maioria dos programas de regressão informa a estatística F em (4.46) automaticamente, o que torna tentador usar essa estatística para testar restrições de exclusão gerais. Você deve evitar essa tentação. A estatística F em (4.41) é usada para restrições de exclusão gerais; ela depende dos R-quadrados dos modelos restrito e irrestrito. A forma especial (4.46) é válida somente para testar a exclusão conjunta de *todas* as variáveis independentes. Às vezes, isso é chamado de teste de **significância geral da regressão**.

Se não for possível rejeitar (4.44), não há evidência de que qualquer uma das variáveis independentes ajude a explicar y. Isso usualmente significa que devemos procurar outras variáveis para explicar y. Para o Exemplo 4.9, a Estatística F para testar (4.44) é cerca de 9,55 com $k = 5$ e $n - k - 1 = 1.185$ *gl*. O p-valor é zero para quatro casas após o ponto decimal, de modo que (4.44) é fortemente rejeitada. Assim, concluímos que as variáveis na equação *pesonas* explicam, *de fato*, alguma variação em *pesonas*. A quantidade explicada não é grande: somente 3,87%. No entanto, o R-quadrado aparentemente pequeno resulta em uma estatística F altamente significante. Essa é a razão de devermos calcular a estatística F para testar a significância conjunta e não apenas olhar o tamanho do R-quadrado.

Ocasionalmente, a estatística F para a hipótese de que todas as variáveis independentes são conjuntamente não significantes pode ser o foco de um estudo. O Problema 4.10 pedirá a você para usar dados de retorno de ações para testar se os retornos das ações ao longo de um horizonte de quatro anos são previsíveis, com base em informações conhecidas somente no início do período. Sob a *hipótese de mercados eficientes*, os retornos não deveriam ser previsíveis; a hipótese nula é precisamente (4.44).

Teste de Restrições Lineares Gerais

Testar restrições de exclusão é, de longe, a mais importante aplicação da estatística F. Às vezes, entretanto, as restrições implicadas por uma teoria são mais complicadas do que apenas excluir algumas variáveis independentes. É ainda simples usar a estatística F para um teste dessa natureza.

Como um exemplo, considere a seguinte equação:

$$\log(preço) = \beta_0 + \beta_1 \log(aval) + \beta_2 \log(tamterr) + \beta_3 \log(arquad) + \beta_4 qtdorm + u, \tag{4.47}$$

em que *preço* é o preço das casas, *aval* é o valor avaliado das casas (antes de elas serem vendidas), *tamterr* é o tamanho dos terrenos, em pés quadrados, *arquad* é a área da casa em pés quadrados e *qtdorm* é o número de quartos. Agora, suponha que gostaríamos de testar se o preço de avaliação das casas é uma avaliação racional. Nesse caso, uma variação de 1% em *aval* deve estar associada a uma variação de 1% em *preço*; isto é, $\beta_1 = 1$. Além disso, *tamterr*, *arquad* e *qtdorm* não devem ajudar a explicar log(*preço*), uma vez que o valor de avaliação tenha sido controlado. Juntas, essas hipóteses podem ser expressas como

$$H_0: \beta_1 = 1, \beta_2 = 0, \beta_3 = 0, \beta_4 = 0. \tag{4.48}$$

Aqui há quatro restrições a serem testadas; três são restrições de exclusão, mas $\beta_1 = 1$ não é. Como podemos testar essa hipótese usando a estatística F?

Como no caso da restrição de exclusão, estimamos o modelo irrestrito, (4.47) nesse caso, e, em seguida, impomos as restrições em (4.48) para obter o modelo restrito. O segundo passo pode ser um pouquinho complicado. Porém, tudo o que fazemos é inserir as restrições. Se escrevermos (4.47) como

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u, \tag{4.49}$$

então o modelo restrito é $y = \beta_0 + x_1 + u$. Agora, a fim de impor a restrição de que o coeficiente de x_1 é a unidade, devemos estimar o seguinte modelo:

$$y - x_1 = \beta_0 + u. \tag{4.50}$$

Esse é apenas um modelo com um intercepto (β_0), mas com uma variável dependente diferente daquela em (4.49). O procedimento para calcular a estatística F é o mesmo: estime (4.50), obtenha o SQR (SQR_r) e utilize-o com o SQR de (4.49) na estatística F (4.37). Estamos testando $q = 4$ restrições, e há $n - 5$ gl no modelo irrestrito. A estatística F é simplesmente $[(SQR_r - SQR_{ir})/SQR_{ir}][(n - 5)/4]$.

Antes de ilustrar esse teste usando um conjunto de dados devemos enfatizar um ponto: não podemos usar a forma R-quadrado da estatística F nesse exemplo porque a variável dependente em (4.50) é diferente daquela em (4.49). Isso significa que a soma dos quadrados total das duas regressões será diferente, e (4.41) não é mais equivalente a (4.37). Como regra geral, a forma SQR da estatística F deve ser usada se uma variável dependente diferente for necessária para computar a regressão restrita.

O modelo irrestrito estimado usando os dados em HPRICE1.RAW é

$$\begin{aligned} \log(\widehat{preço}) &= 0{,}264 + 1{,}043\log(aval) + 0{,}0074\log(tamterr) \\ &\quad (0{,}570)\ (0{,}151)\qquad\quad (0{,}0386) \\ &\quad - 0{,}1032\log(arquad) + 0{,}0338\, qtdorm \\ &\quad (0{,}1384)\qquad\quad (0{,}0221) \\ n &= 88,\ SQR = 1{,}822,\ R^2 = 0{,}773. \end{aligned}$$

Se usamos separadamente as estatísticas t para testar cada hipótese em (4.48), não será possível rejeitar cada uma delas. Contudo, a racionalidade da avaliação é uma hipótese conjunta, de modo que devemos testar as restrições conjuntamente. O SQR do modelo restrito é $SQR_r = 1{,}880$ e, portanto, a estatística F é $[(1{,}880 - 1{,}822)/1{,}822](83/4) = 0{,}661$. O valor crítico de 5% em uma distribuição F com (4,83) gl é cerca de 2,50, e portanto não podemos rejeitar H_0. Não há, essencialmente, evidência contra a hipótese de que os valores de avaliação sejam racionais.

4.6 DESCRIÇÃO DOS RESULTADOS DA REGRESSÃO

Finalizamos este capítulo dando algumas instruções de como descrever os resultados da regressão múltipla para projetos empíricos relativamente complicados. Isso deve ensiná-lo a ler trabalhos publicados nas ciências sociais aplicadas, ao mesmo tempo em que prepara você para escrever seus próprios artigos empíricos. Expandiremos este tópico no restante do livro ao descrever os resultados de vários exemplos, mas muitos dos pontos fundamentais podem ser apresentados agora.

Naturalmente, os coeficientes estimados de MQO devem ser sempre informados. Das variáveis fundamentais de uma análise, você deve *interpretar* os coeficientes estimados (o que, freqüentemente, requer conhecer as unidades de medida das variáveis). Por exemplo, ele é uma elasticidade ou tem alguma outra interpretação que necessita de explicação? A importância econômica ou prática das estimativas das variáveis-chave devem ser discutidas.

Os erros-padrão devem sempre ser incluídos juntamente com os coeficientes estimados. Alguns autores preferem informar as estatísticas t em vez dos erros-padrão (e, freqüentemente, apenas o valor absoluto das estatísticas t). Embora não haja, realmente, nada de errado com isso, há alguma preferên-

cia por informar os erros-padrão. Primeiro, isso nos força a pensar, cuidadosamente, sobre a hipótese nula que está sendo testada; a hipótese nula nem sempre corresponde a dizer que o parâmetro populacional é zero. Segundo, ter os erros-padrão torna mais fácil calcular os intervalos de confiança.

O R-quadrado da regressão deve sempre ser incluído. Vimos que, além de dar uma medida do grau de ajuste, ele faz com que os cálculos das estatísticas F para as restrições de exclusão fiquem simples. Informar a soma dos resíduos quadrados e o erro-padrão da regressão às vezes é uma boa idéia, mas não é crucial. O número de observações usado na estimação de qualquer equação deve aparecer próximo da equação estimada.

Se somente alguns poucos modelos são estimados, os resultados podem ser resumidos na forma de equações, como fizemos até aqui. Entretanto, em muitos trabalhos, várias equações são estimadas para muitos conjuntos diferentes de variáveis independentes. Podemos estimar a mesma equação para diferentes grupos de pessoas, ou mesmo ter equações que explicam diferentes variáveis dependentes. Em tais casos, é melhor resumir os resultados em uma ou mais tabelas. A variável dependente deve ser indicada claramente na tabela, e as variáveis independentes, listadas na primeira coluna. Os erros-padrão (ou as estatísticas t) podem ser colocados em parênteses abaixo das estimativas.

EXEMPLO 4.10

(A Relação Salário-Benefícios de Professores)

Façamos *totrem* representar a remuneração média anual total de um professor, incluindo o salário e todos os benefícios adicionais (pensão, seguro-saúde etc.). Ampliando a equação dos salários, a remuneração total deve ser uma função da produtividade e talvez de outras características. Como é padrão, vamos usar a forma logarítmica:

$$\log(totrem) = f(características\ da\ produtividade, outros\ fatores),$$

em que $f(\cdot)$ é alguma função (não-especificada por enquanto). Escreva

$$totrem = salário + benefícios = salário\left(1 + \frac{\text{benefícios}}{\text{salário}}\right).$$

Essa equação mostra que a remuneração total é o produto de dois termos: *salário* e 1 + *b/s*, em que *b/s* é a abreviação para "razão benefícios-salário". Tirando o log dessa equação resulta em log(*totrem*) = log(*salário*) + log(1 + *b/s*). Agora, para um *b/s* "pequeno", log(1 + *b/s*) ≈ *b/s*; vamos usar essa aproximação. Isso leva ao modelo econométrico

$$\log(salário) = \beta_0 + \beta_1(b/s) + outros\ fatores.$$

Testar a relação salário-benefícios é, então, o mesmo que testar H_0: $\beta_1 = -1$ contra H_1: $\beta_1 \neq -1$.

Vamos usar os dados do arquivo MEAP93.RAW para testar essa hipótese. Esses dados são ponderados por escola, e não observamos muitos outros fatores que poderiam afetar a remuneração total. Incluiremos os controles para o tamanho da escola (*matricl*), número de funcionários por mil estudantes e medidas como taxas de evasão escolar (*taxevas*) e de formatura (*taxform*). O *b/s* médio na amostra é cerca de 0,205, e o maior valor é 0,450.

As equações estimadas são apresentadas na Tabela 4.1, na qual os erros-padrão aparecem entre parênteses, abaixo das estimativas dos coeficientes. A variável-chave é *b/s*, a razão benefícios-salário.

EXEMPLO 4.10 (continuação)

Na primeira coluna da Tabela 4.1, vemos que, sem controlar quaisquer outros fatores, o coeficiente de MQO de *b/s* é −0,825. A estatística *t* para testar a hipótese nula H_0: $\beta_1 = -1$ é $t = (-0,825 + 1)/0,200 = 0,875$, e portanto a regressão simples não permite rejeitar H_0. Após adicionar controles para o tamanho da escola e o tamanho do corpo docente (o qual captura, mais ou menos, o número de estudantes por professor), a estimativa do coeficiente de *b/s* passa a ser −0,605. Agora, o teste de $\beta_1 = -1$ resulta em uma estatística *t* igual a cerca de 2,39; assim, H_0 é rejeitada ao nível de 5% contra uma alternativa bilateral. As variáveis log(*matricl*) e log(*staff*) são estatisticamente muito significantes.

De que modo o acréscimo de *taxevas* e *taxform* afeta a estimativa da *relação* salário-benefícios? Essas variáveis são conjuntamente significantes ao nível de 5%? E ao nível de 10%?

Tabela 4.1

Teste da Relação Salário-Benefícios

Variável Dependente: log(*salário*)			
Variáveis Independentes	**(1)**	**(2)**	**(3)**
b/s	−0,825 (0,200)	−0,605 (0,165)	−0,589 (0,165)
log(*matricl*)	—	0,874 (0,0073)	0,0881 (0,0073)
log(*staff*)	—	−0,222 (0,050)	−0,218 (0,050)
taxevas	—	—	−0,00028 (0,00161)
taxfor	—	—	0,00097 (0,00066)
intercepto	10,523 (0,042)	10,884 (0,252)	10,738 (0,258)
Observações *R*-quadrado	408 0,040	408 0,353	408 0,361

Neste capítulo, cobrimos o tópico muito importante da inferência estatística, o qual nos permite obter conclusões sobre o modelo populacional a partir de uma amostra aleatória. Vamos resumir os pontos principais:

1. Sob as hipóteses do modelo linear clássico RLM.1 a RLM.6, os estimadores de MQO são normalmente distribuídos.
2. Sob as hipóteses do MLC, as estatísticas t têm distribuições t sob a hipótese nula.
3. Usamos as estatísticas t para testar hipóteses sobre um único parâmetro contra alternativas unilaterais ou bilaterais, usando testes monocaudais ou bicaudais, respectivamente. A hipótese nula mais comum é H_0: $\beta_j = 0$, mas, às vezes, queremos testar outros valores de β_j sob H_0.
4. No teste de hipótese clássico, primeiro escolhemos um nível de significância que, juntamente com os gl e a hipótese alternativa, determina o valor crítico contra o qual comparamos a estatística t. É mais informativo calcular o p-valor de um teste t – o nível de significância menor ao qual a hipótese nula é rejeitada –, de modo que a hipótese pode ser testada a qualquer nível de significância.
5. Sob as hipóteses do MLC, os intervalos de confiança podem ser construídos para cada β_j. Esses ICs podem ser usados para testar qualquer hipótese nula relativa a β_j contra uma alternativa bilateral.
6. Testes de hipóteses simples relativos a mais de um β_j podem sempre ser testados, reescrevendo o modelo de tal forma que ele contenha o parâmetro de interesse. Em seguida, uma estatística t padrão pode ser usada.
7. A estatística F é usada para testar restrições múltiplas de exclusão, e há duas formas equivalentes do teste. Uma está baseada nos SQRs dos modelos restrito e irrestrito. Uma forma mais conveniente está baseada nos R-quadrados dos dois modelos.
8. Ao calcular uma estatística F, os gl do numerador correspondem ao número de restrições que estão sendo testadas, enquanto os gl do denominador são os graus de liberdade do modelo irrestrito.
9. A hipótese alternativa do teste F é bilateral. Na abordagem clássica especificamos um nível de significância que, juntamente com o gl do numerador e o gl do denominador, determina o valor crítico. A hipótese nula é rejeitada quando a estatística F excede o valor crítico c. Alternativamente, podemos calcular o p-valor para resumir a evidência contra H_0.
10. Restrições lineares múltiplas gerais podem ser testadas usando a forma soma dos resíduos quadrados da estatística F.
11. A estatística F da significância geral de uma regressão testa a hipótese nula de que *todos* os parâmetros de inclinação são zero, com o intercepto irrestrito. Sob H_0, as variáveis explicativas não têm efeito sobre o valor esperado de y.

4.1 Quais dos seguintes itens podem fazer com que as estatísticas de MQO não sejam válidas (isto é, que elas não tenham distribuições t sob H_0)?

(i) Heteroscedasticidade.
(ii) Um coeficiente de correlação de 0,95 entre duas variáveis independentes que estão no modelo.
(iii) Omitir uma variável explicativa importante.

4.2 Considere uma equação para explicar os salários dos diretores executivos em termos das vendas anuais das empresas (*vendas*), dos retornos das ações sobre o patrimônio (*rma*, na forma percentual) e dos retornos das ações sobre o valor das ações das empresas (*raf*, na forma percentual):

$$\log(salário) = \beta_0 + \beta_1 \log(vendas) + \beta_2 \log(rma) + \beta_3 \log(raf) + u.$$

(i) Em termos dos parâmetros do modelo, formule a hipótese nula em que, após controlar *vendas* e *rma*, *raf* não tem efeito sobre o salário dos diretores executivos. Formule a hipótese alternativa de que um melhor desempenho de mercado das ações aumenta o salário dos diretores executivos.

(ii) Usando os dados em CEOSAL1.RAW, obteve-se a seguinte equação por MQO:

$$\widehat{\log(salário)} = 4{,}32 + 0{,}280 \log(vendas) + 0{,}0174\, rma + 0{,}00024\, raf$$
$$(0{,}32)\ \ (0{,}035)\qquad\qquad (0{,}0041)\qquad (0{,}00054)$$
$$n = 209,\ R^2 = 0{,}283.$$

Se *raf* aumenta em 50 pontos, qual é a variação percentual prevista em *salário*? Na prática, *raf* tem um efeito grande sobre *salário*?

(iii) Teste a hipótese nula de que *raf* não tem efeito sobre *salário* contra a alternativa de que *raf* tem um efeito positivo. Faça o teste ao nível de significância de 10%.

(iv) Você incluiria *raf* no modelo final que explica a remuneração dos diretores executivos em termos do desempenho das empresas? Explique.

4.3 A variável *pdintens* corresponde a gastos com pesquisa e desenvolvimento (P&D) como uma percentagem das vendas. As vendas são mensuradas em milhões de dólares. A variável *lucrmarg* corresponde a lucros como uma percentagem das vendas.

Usando os dados do arquivo RDCHEM.RAW de 32 empresas da indústria química, estimou-se a seguinte equação:

$$\widehat{pdintens} = 0{,}472 + 0{,}321 \log(vendas) + 0{,}050\, lucrmarg$$
$$(1{,}369)\ \ (0{,}216)\qquad\qquad (0{,}046)$$
$$n = 32,\ R^2 = 0{,}099.$$

(i) Interprete o coeficiente de log(*vendas*). Em particular, se *vendas* aumenta em 10%, qual é a variação percentual estimada em *pdintens*? Esse efeito é economicamente grande?

(ii) Teste a hipótese de que a intensidade de P&D não varia com *vendas* contra a alternativa de que P&D aumenta com as vendas. Teste aos níveis de 5% e 10%.

(iii) *lucrmarg* tem um efeito estatisticamente significante sobre *pdintens*?

4.4 As taxas de aluguel são influenciadas pela população de estudantes em uma cidade onde há universidades? Seja *alug* o aluguel médio mensal pago pela unidade alugada em uma cidade nos Estados Unidos, onde há universidades. Seja *pop* o total da população da cidade, *rendmed*, a renda média da cidade e *pctestu*, a população de estudantes como um percentual da população total. Um modelo para testar uma relação é

$$\log(alug) = \beta_0 + \beta_1 \log(pop) + \beta_2 \log(rendmed) + \beta_3\, pctestu + u.$$

(i) Formule a hipótese nula de que o tamanho da população estudantil relativo à população das cidades não tem efeito *ceteris paribus* sobre os aluguéis mensais. Formule a alternativa de que há um efeito.

(ii) Quais sinais você espera para β_1 e β_2?

(iii) A equação estimada, usando 1.990 dados de 64 cidades com universidades do arquivo RENTAL.RAW, é

$$\widehat{\log(alug)} = 0{,}043 + 0{,}066\log(pop) + 0{,}507\log(rendmed) + 0{,}0056\,pctestu$$
$$(0{,}844)\quad (0{,}039)\qquad\qquad (0{,}081)\qquad\qquad (0{,}0017)$$
$$n = 64,\ R^2 = 0{,}458.$$

O que está errado com a seguinte afirmação: "Um aumento de 10% na população está associado a um aumento de cerca de 6,6% no aluguel"?

(iv) Teste a hipótese formulada na parte (i) ao nível de 1%.

4.5 Considere a equação estimada do Exemplo 4.3 (que poderia também ser usada para estudar os efeitos de faltar às aulas sobre a nota média em curso superior):

$$\widehat{nmgrad} = 1{,}39 + 0{,}412\,nmem + 0{,}15\,tac - 0{,}083\,faltas$$
$$(0{,}33)\quad (0{,}94)\qquad\qquad (0{,}011)\qquad (0{,}026)$$
$$n = 141,\ R^2 = 0{,}234.$$

(i) Usando a aproximação normal padronizada, encontre o intervalo de confiança de 95% para β_{nmem}.

(ii) Você pode rejeitar a hipótese H_0: $\beta_{nmem} = 0{,}4$ contra a hipótese alternativa bilateral ao nível de 5%?

(iii) Você pode rejeitar a hipótese H_0: $\beta_{nmem} = 1$ contra a hipótese alternativa bilateral ao nível de 5%?

4.6 Na Seção 4.5 usamos, como exemplo, o teste da racionalidade da avaliação dos preços de casas. Lá, usamos um modelo log-log em *preço* e *aval* [veja equação (4.47)]. Aqui, vamos usar uma formulação nível-nível.

(i) No modelo de regressão simples

$$preço = \beta_0 + \beta_1 aval + u,$$

a avaliação é racional se $\beta_1 = 1$ e $\beta_0 = 0$. A equação estimada é

$$\widehat{prêço} = -14{,}47 + 0{,}976\,aval$$
$$(16{,}27)\quad (0{,}049)$$
$$n = 88,\ SQR = 165{,}644.51,\ R^2 = 0{,}820.$$

Primeiro, teste a hipótese H_0: $\beta_0 = 0$ contra a hipótese alternativa bilateral. Em seguida, teste H_0: $\beta_1 = 1$ contra a hipótese alternativa bilateral. O que você conclui?

(ii) Para testar a hipótese conjunta $\beta_0 = 0$ e $\beta_1 = 1$, precisamos do SQR do modelo restrito. Isso é igual a calcular $\sum_{i=1}^{n} (preço_i - aval_i)^2$, em que $n = 88$, visto que os resíduos do modelo

restrito são exatamente $preço_i - aval_i$. (Nenhuma estimação é necessária para o modelo restrito porque ambos os parâmetros estão especificados sob H_0.) Isso tem como resultado SQR = 209.448,99. Faça o teste F para a hipótese conjunta.

(iii) Agora, teste H_0: $\beta_2 = 0, \beta_3 = 0$ e $\beta_4 = 0$ no modelo

$$preço = \beta_0 + \beta_1 aval + \beta_2 tamterr + \beta_3 arquad + \beta_4 qtdorm + u.$$

O R-quadrado da estimação desse modelo usando as mesmas 88 residências é 0,829.

(iv) Se a variância de *preço* varia com *aval*, *tamterr*, *arquad* ou *qtdorm*, o que você pode dizer sobre o teste F da parte (iii)?

4.7 No Exemplo 4.7, usamos dados de empresas manufatureiras de Michigan para estimar a relação entre a taxa de rejeição e outras características da firma. Agora, vamos olhar esse exemplo mais de perto e usar uma amostra maior de empresas.

(i) O modelo populacional estimado no Exemplo 4.7 pode ser escrito como

$$\log(rejei) = \beta_0 + \beta_1 hrsemp + \beta_2 \log(vendas) + \beta_3 \log(empreg) + u.$$

Usando as 43 observações disponíveis para 1987, a equação estimada é

$$\log(\widehat{rejei}) = 11{,}74 - 0{,}042\ hrsemp - 0{,}951\ \log(vendas) + 0{,}992\ \log(empreg)$$
$$(4{,}57)\ \ (0{,}019)\qquad\qquad (0{,}370)\qquad\qquad (0{,}360)$$
$$n = 43,\ R^2 = 0{,}310.$$

Compare essa equação com aquela estimada com somente 30 firmas na amostra.

(ii) Mostre que o modelo populacional pode também ser escrito como

$$\log(\widehat{rejei}) = \beta_0 + \beta_1 \log(hrsemp) + \beta_2 \log(vendas/empreg) + \theta_3 \log(empreg) + u.$$

em que $\theta_3 \equiv \beta_2 + \beta_3$. [*Sugestão*: Lembre-se de que $\log(x_2/x_3) = \log(x_2) - \log(x_3)$.]

Interprete a hipótese H_0: $\theta_3 = 0$.

(iii) Quando a equação da parte (ii) é estimada, obtemos

$$\log(\widehat{rejei}) = 11{,}74 - 0{,}042\ hrsemp - 0{,}951\ \log(vendas/empreg) + 0{,}041\ \log(empreg)$$
$$(4{,}57)\ \ (0{,}019)\qquad\qquad (0{,}370)\qquad\qquad (0{,}205)$$
$$n = 43,\ R^2 = 0{,}310.$$

Controlando o treinamento dos trabalhadores e a razão vendas-empregados, as empresas maiores têm taxas de rejeição maiores estatisticamente significantes?

(iv) Teste a hipótese de que um aumento de 1% em *vendas/empreg* está associado a uma queda de 1% na taxa de rejeição.

4.8 Considere o modelo de regressão múltipla com três variáveis independentes, sob as hipóteses do modelo linear clássico RLM.1 a RLM.6:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

Você deseja testar a hipótese nula H_0: $\beta_1 - 3\beta_2 = 1$.

(i) Sejam $\hat{\beta}_1$ e $\hat{\beta}_2$ os estimadores de MQO de β_1 e β_2. Encontre $\text{Var}(\hat{\beta}_1 - 3\hat{\beta}_2)$ em termos das variâncias de $\hat{\beta}_1$ e $\hat{\beta}_2$ e a covariância entre eles. Qual é o erro-padrão de $\hat{\beta}_1 - 3\hat{\beta}_2$?

(ii) Escreva a estatística t para testar H_0: $\beta_1 - 3\beta_2 = 1$.

(iii) Defina $\theta_1 = \beta_1 - 3\beta_2$ e $\hat{\theta}_1 = \hat{\beta}_1 - 3\hat{\beta}_2$. Escreva uma equação de regressão que envolva β_0, θ_1, β_2 e β_3, que permita que você obtenha diretamente $\hat{\theta}_1$ e seu erro-padrão.

4.9 No Problema 3.3, estimamos a equação

$$\widehat{dormir} = 3.638{,}25 - 0{,}148\ trabtot - 11{,}13\ educ + 2{,}20\ idade$$
$$(112{,}28)\ \ (0{,}017)\qquad\quad (5{,}88)\qquad\ (1{,}45)$$
$$n = 706,\ R^2 = 0{,}113,$$

para a qual informamos, agora, os erros-padrão juntamente com as estimativas.

(i) *educ* ou *idade* são individualmente significantes ao nível de 5% contra uma hipótese alternativa bilateral? Mostre como você chegou à resposta.

(ii) Ao retirar *educ* e *idade* da equação, temos

$$\widehat{dormir} = 3.586{,}25 - 0{,}151\ trabtot$$
$$(38{,}91)\ \ (0{,}017)$$
$$n = 706,\ R^2 = 0{,}103,$$

É possível afirmar que *educ* e *idade* são conjuntamente significantes na equação original ao nível de 5%? Justifique sua resposta.

(iii) Incluir *educ* e *idade* no modelo afeta muito a relação estimada entre dormir e trabalhar?

(iv) Suponha que a equação de dormir contenha heteroscedasticidade. O que isso significa para os testes calculados nas partes (i) e (ii)?

4.10 A análise de regressão pode ser usada para testar se o mercado usa eficientemente as informações ao avaliar ações. Seja *retorno* o retorno total de manter ações de uma firma ao longo de um período de quatro anos, do final de 1990 até o final de 1994. A *hipótese de mercados eficientes* diz que esses retornos não devem estar sistematicamente relacionados à informação conhecida em 1990. Se as características conhecidas no início do período ajudarem a prever os retornos das ações, poderíamos usar essas informações para escolher as ações.

Para 1990, seja *rdc* a relação dívida-capital de uma empresa, seja *gpa* os ganhos por ação, seja *rendliq* a renda líquida e seja *salário* a remuneração total dos diretores executivos da empresa.

(i) Usando os dados do arquivo RETURN.RAW, estimou-se a seguinte equação:

$$\widehat{retorno} = -14{,}37 + 0{,}321\ rdc + 0{,}043\ gpa - 0{,}0051\ rendliq + 0{,}0035\ salário$$
$$(6{,}89)\ \ (0{,}201)\qquad (0{,}078)\qquad\ (0{,}0047)\qquad\ (0{,}0022)$$
$$n = 142,\ R^2 = 0{,}0395.$$

Teste se as variáveis explicativas são conjuntamente significantes ao nível de 5%? Alguma variável explicativa é individualmente significante?

(ii) Agora, nova estimação do modelo usando a forma log para *rendliq* e *salário* forneceu a seguinte equação:

$$\widehat{retorno} = -36{,}30 + 0{,}327\, rdc + 0{,}069\, gpa - 4{,}74 \log(rendliq) + 7{,}24 \log(salário)$$
$$(39{,}37)\quad (0{,}203)\qquad (0{,}080)\qquad (3{,}39)\qquad\qquad (6{,}31)$$
$$n = 142,\ R^2 = 0{,}0330.$$

Alguma de suas conclusões da parte (i) mudou?

(iii) Por que não usamos também os logs de *rdc* e *gpa* na parte (ii)?

(iv) Em geral, a evidência da previsibilidade dos retornos é forte ou fraca?

4.11 A tabela seguinte foi criada ao usar os dados do arquivo CEOSAL2. RAW:

Variável Dependente: log(*salário*)			
Variáveis Independentes	**(1)**	**(2)**	**(3)**
log(*vendas*)	0,224 (0,027)	0,158 (0,040)	0,188 (0,040)
log(*valmerc*)	—	0,112 (0,050)	0,100 (0,049)
lucrmarg	—	−0,0023 (0,0022)	−0,0022 (0,0021)
permceo	—	—	0,0171 (0,0055)
percomp	—	—	−0,0092 (0,0033)
intercepto	4,94 (0,20)	4,62 (0,25)	4,57 (0,25)
Observações	177	177	177
R-quadrado	0,281	0,304	0,353

A variável *valmerc* é o valor de mercado da firma, *lucrmarg* é o lucro como percentagem das vendas, *permceo* corresponde aos anos trabalhando como diretor executivo na atual companhia e *percomp* é o total de anos na companhia.

(i) Comente os efeitos de *lucrmarg* sobre o salário dos diretores executivos.

(ii) O valor de mercado tem um efeito significante? Explique.

(iii) Interprete os coeficientes de *permceo* e *percomp*. As variáveis são estatisticamente significantes?

(iv) O que você entende do fato de que a permanência muito longa na companhia, mantendo fixos os outros fatores, está associada a salários mais baixos?

Análise de Regressão Múltipla: MQO Assimptótico

os capítulos 3 e 4 estudamos o que chamamos propriedades de *amostra finita*, de *amostra pequena* ou *exatas* dos estimadores de MQO do modelo populacional

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u . \qquad \textbf{(5.1)}$$

Por exemplo, a inexistência de viés de MQO (derivada no Capítulo 3), sob as quatro primeiras hipóteses de Gauss-Markov, é uma propriedade de amostra finita porque ela é válida para *qualquer* amostra de tamanho n (sujeita à restrição amena de que n deve ser pelo menos tão grande quanto o número total de parâmetros no modelo de regressão, $k + 1$). Semelhantemente, o fato de que MQO é o melhor estimador não-viesado linear sob o conjunto completo das hipóteses de Gauss-Markov (RLM.1 a RLM.6) é uma propriedade de amostra finita.

No Capítulo 4 acrescentamos a hipótese do modelo linear clássico RLM.6, a qual afirma que o termo erro u é normalmente distribuído e independente das variáveis explicativas. Isso nos permitiu derivar as distribuições amostrais *exatas* dos estimadores de MQO (condicionados às variáveis explicativas da amostra). Em particular, o Teorema 4.1 mostrou que os estimadores de MQO têm distribuições amostrais normais, o que levou diretamente às distribuições t e F das estatísticas t e F. Se o erro não é normalmente distribuído, a distribuição de uma estatística t não é exatamente t, e uma estatística F não tem uma distribuição F exata para qualquer tamanho de amostra.

Além das propriedades de amostra finita é importante conhecer as **propriedades assimptóticas** ou **propriedades de amostras grandes** dos estimadores e das estatísticas de testes. Essas propriedades não são definidas para um tamanho particular de amostra; pelo contrário, elas são definidas quando o tamanho da amostra cresce sem limites. Felizmente, sob as hipóteses que fizemos, o método MQO tem propriedades de amostra grande satisfatórias. Uma constatação importante na prática é que mesmo sem a hipótese de normalidade (hipótese RLM.6), as estatísticas t e F têm distribuições *aproximadamente* t e F, pelo menos em amostras grandes. Vamos discutir esse assunto com mais detalhes na Seção 5.2, após compreendermos a consistência do método MQO na Seção 5.1.

5.1 CONSISTÊNCIA

A inexistência de viés dos estimadores, embora importante, não pode ser conseguida sempre. Por exemplo, como discutimos no Capítulo 3, o erro-padrão da regressão, $\hat{\sigma}$, não é um estimador não-viesado de σ, o desvio-padrão do erro u em um modelo de regressão múltipla.

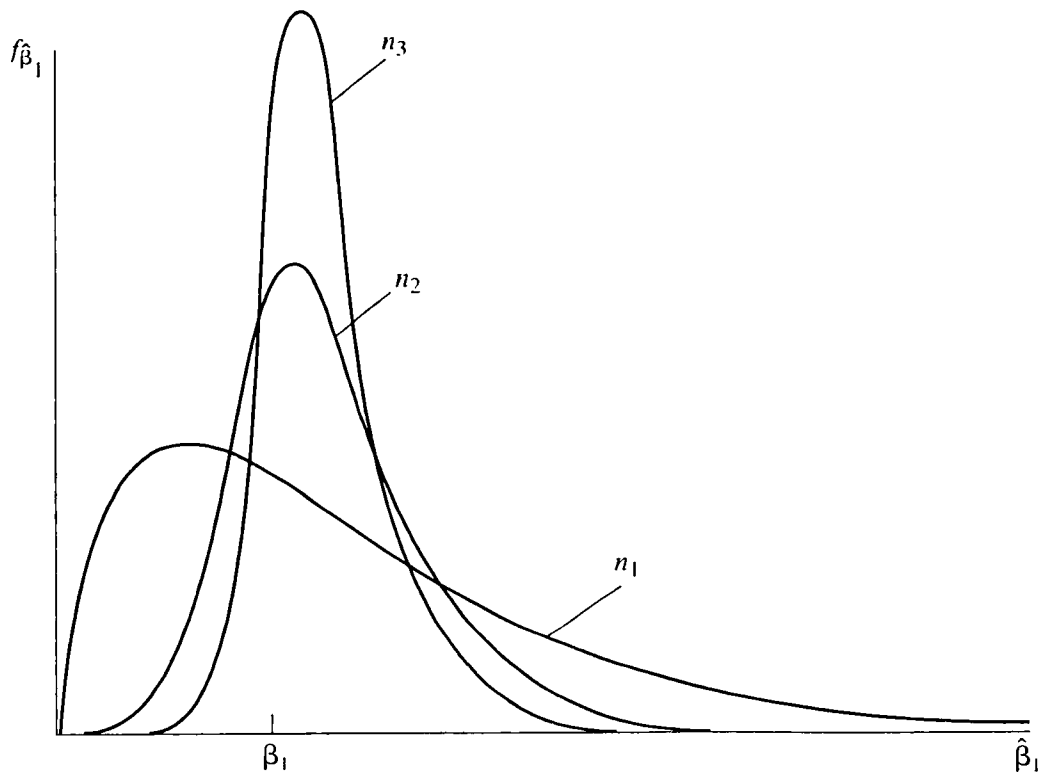
Embora os estimadores de MQO sejam não-viesados sob RLM.1 a RLM.4, descobriremos, no Capítulo 11, que há regressões de séries de tempo em que os estimadores de MQO não são não-viesados. Além disso, na Parte 3 do livro, encontraremos vários outros estimadores que são viesados.

Embora nem todos os estimadores úteis sejam não-viesados, virtualmente todos os economistas concordam que a **consistência** é um requisito mínimo de um estimador. O famoso econometrista Clive W. J. Granger observou certa vez: "Se você não puder obter a consistência apropriadamente quando n tende ao infinito, você não deveria se envolver com isso". A implicação é que, quando seu estimador de um parâmetro populacional particular não for consistente, você estará desperdiçando seu tempo.

Há maneiras um pouco diferentes de descrever a consistência. As definições e os resultados formais estão apresentados no Apêndice C, disponível em www.thomsonlearning.com.br, aqui vamos dar ênfase a um entendimento intuitivo. Mais concretamente, seja $\hat{\beta}_j$ o estimador de MQO de β_j para algum j. Para cada n, $\hat{\beta}_j$ tem uma distribuição de probabilidades (representando seus valores possíveis em diferentes amostras aleatórias de tamanho n). Como $\hat{\beta}_j$ é não-viesado sob as hipóteses RLM.1 a RLM.4, essa distribuição tem valor médio β_j. Se esse estimador for consistente, a distribuição de $\hat{\beta}_j$ se torna mais e mais estreitamente distribuída ao redor de β_j quando o tamanho da amostra cresce. Quando n tende ao infinito, a distribuição de $\hat{\beta}_j$ encontra-se no ponto único β_j. De fato, isso significa que podemos fazer com que nosso estimador, arbitrariamente, aproxime-se de β_j se pudermos coletar tantos dados quanto desejarmos. Essa convergência está ilustrada na Figura 5.1.

Figura 5.1

Distribuições amostrais de $\hat{\beta}_1$ para amostras de tamanhos $n_1 < n_2 < n_3$.



Naturalmente, para qualquer aplicação, temos um tamanho de amostra fixo, que é a razão pela qual uma propriedade assimptótica tal como a consistência pode ser difícil de entender. **Consistência** envolve um experimento mental sobre o que aconteceria se o tamanho da amostra se tornasse grande (enquanto, ao mesmo tempo, obtemos muitas amostras aleatórias para cada tamanho de amostra). Se a obtenção de mais e mais dados não nos levar, em geral, para perto do valor do parâmetro de interesse, estamos usando um procedimento de estimação insatisfatório.

Convenientemente, o mesmo conjunto de hipóteses implica tanto a inexistência de viés como a consistência de MQO. Vamos resumir o assunto com um teorema.

TEOREMA 5.1 (CONSISTÊNCIA DE MQO)

Sob as hipóteses RLM.1 a RLM.4, o estimador de MQO $\hat{\beta}_j$ é um estimador consistente de β_j, para todo $j = 0, 1, \ldots, k$.

Uma prova geral desse resultado é mais facilmente desenvolvida usando os métodos da álgebra matricial descritos nos Apêndices D e E, disponíveis na página do livro, no site www.thomsonlearning.com.br. No entanto, podemos provar o Teorema 5.1 sem dificuldades no caso do modelo de regressão simples. Vamos nos concentrar no estimador de inclinação, $\hat{\beta}_1$.

A prova começa do mesmo modo que a prova da inexistência de viés: escrevemos a fórmula de $\hat{\beta}_1$, e em seguida a inserimos em $y_i = \beta_0 + \beta_1 x_{i1} + u_i$:

$$\begin{aligned}\hat{\beta}_1 &= \left(\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)y_i\right)\Big/\left(\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2\right) \\ &= \beta_1 + \left(n^{-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)u_i\right)\Big/\left(n^{-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2\right).\end{aligned} \tag{5.2}$$

Podemos aplicar a lei dos grandes números ao numerador e ao denominador, os quais convergem em probabilidade para as quantidades populacionais, $\text{Cov}(x_1,u)$ e $\text{Var}(x_1)$, respectivamente. Como $\text{Var}(x_1) \neq 0$ — o que é assumido em RLM.4 — podemos usar as propriedades dos *limites de probabilidade* (veja o Apêndice C, disponível no site da Thomson) para obter

$$\begin{aligned}\text{plim}\,\hat{\beta}_1 &= \beta_1 + \text{Cov}(x_1, u)/\text{Var}(x_1) \\ &= \beta_1, \text{ porque } \text{Cov}(x_1, u) = 0.\end{aligned} \tag{5.3}$$

Usamos o fato, discutido nos capítulos 2 e 3, de que $E(u|1) = 0$ implica que x_1 e u são não-correlacionados (têm covariância zero).

Como questão técnica, para garantir que os limites de probabilidade existam, devemos assumir que $\text{Var}(x_1) < \infty$ e $\text{Var}(u) < \infty$ (o que significa que suas distribuições de probabilidade não são muito espalhadas), mas não nos preocuparemos com casos em que essas hipóteses não se mantenham.

Os argumentos anteriores, e a equação (5.3) em particular, mostram que MQO é consistente no caso da regressão simples se assumirmos somente correlação zero. Isso também é verdadeiro no caso geral. Vamos agora formular isso como uma hipótese.

HIPÓTESE RLM.3' (MÉDIA ZERO E CORRELAÇÃO ZERO)
$E(u) = 0$ e $Cov(x_j,u) = 0$, para $j = 1, 2, ..., k$.

No Capítulo 3 discutimos por que a hipótese RLM.3 implica RLM.3', mas não vice-versa. O fato de que MQO é consistente sob a hipótese mais fraca RLM.3' se revelará útil no Capítulo 15 e em outras situações. É interessante observar que, embora MQO seja não-viesado sob RLM.3, esse não é o caso sob a hipótese RLM.3'. (Essa é a razão primeira de termos assumido RLM.3.)

A Derivação da Inconsistência no Método MQO

Do mesmo modo que a não-observância de $E(u|x_1, ..., x_k) = 0$ causa viés dos estimadores de MQO, a correlação entre u e *qualquer* das variáveis $x_1, x_2, ..., x_k$ faz com que, em geral, *todos* os estimadores de MQO sejam inconsistentes. Essa simples, mas importante, observação é freqüentemente resumida como: *se o erro é correlacionado com qualquer uma das variáveis independentes, MQO é viesado e inconsistente.* Isso representa muita falta de sorte, porque significa que qualquer viés persiste quando o tamanho da amostra cresce.

No caso da regressão simples, podemos obter a inconsistência da primeira parte da equação (5.3), que se mantém sejam u e x_1 não-correlacionados ou não. A **inconsistência** em $\hat{\beta}_1$ (às vezes, imprecisamente chamada de **viés assimptótico**) é

$$\text{plim}\,\hat{\beta}_1 - \beta_1 = \text{Cov}(x_1,u)/\text{Var}(x_1). \qquad \textbf{(5.4)}$$

Como $\text{Var}(x_1) > 0$, a inconsistência em $\hat{\beta}_1$ é positiva se x_1 e u são positivamente correlacionados, e a inconsistência é negativa se x_1 e u são negativamente correlacionados. Se a covariância entre x_1 e u é pequena relativamente à variância em x_1, a inconsistência pode ser desprezível; infelizmente, não podemos nem mesmo estimar quão grande é a covariância porque u não é observado.

Podemos usar (5.4) para derivar o análogo assimptótico do viés de variável omitida (veja Tabela 3.2 no Capítulo 3). Suponha que o modelo verdadeiro,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v,$$

satisfaça as quatro primeiras hipóteses de Gauss-Markov. Então, v tem média zero e é não-correlacionado com x_1 e x_2. Se $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$ foram os estimadores de MQO da regressão de y sobre x_1 e x_2, o Teorema 5.1 implica que esses estimadores são consistentes. Se omitirmos x_2 da regressão e fizermos a regressão simples de y sobre x_1, $u = \beta_2 x_2 + v$. Seja $\tilde{\beta}_1$ o estimador de inclinação da regressão simples. Então

$$\text{plim}\,\tilde{\beta}_1 = \beta_1 + \beta_2\delta_1 \qquad \textbf{(5.5)}$$

em que

$$\delta_1 = \text{Cov}(x_1,x_2)/\text{Var}(x_1). \qquad \textbf{(5.6)}$$

Assim, para propósitos práticos, podemos ver a inconsistência como idêntica ao viés. A diferença é que a inconsistência é expressa em termos da variância populacional de x_1 e da covariância populacional

entre x_1 e x_2, enquanto o viés é baseado em suas contrapartes amostrais (porque estabelecemos condicionamento aos valores de x_1 e x_2 na amostra).

Se x_1 e x_2 forem não-correlacionados (na população), então $\delta_1 = 0$ e $\tilde{\beta}_1$ é um estimador consistente de β_1 (embora não necessariamente não-viesado). Se x_2 tiver um efeito parcial positivo sobre y, de modo que $\beta_2 > 0$, e x_1 e x_2 forem positivamente correlacionados, de modo que $\delta_1 > 0$, a inconsistência em $\tilde{\beta}_1$ é positiva, e assim por diante. Podemos obter a direção da inconsistência ou o viés assimptótico a partir da Tabela 3.2. Se a covariância entre x_1 e x_2 for pequena relativamente à variância de x_1, a inconsistência pode ser pequena.

EXEMPLO 5.1

(Preços de Casas e Distância de um Incinerador)

Seja y o preço de uma casa (*preço*), x_1 a distância da casa a um novo incinerador de lixo (*distância*), e x_2 a "qualidade" da casa (*qualidade*). A variável *qualidade* é imprecisa, de modo que ela pode incluir coisas como o tamanho da casa e do terreno, número de quartos e de banheiros, e, intangíveis, coisas como a atratividade da vizinhança. Se o incinerador deprecia os preços das casas, então β_1 deve ser positivo: tudo mais sendo igual, uma casa que está mais distante do incinerador é mais valiosa. Por definição, β_2 é positivo, visto que casas de qualidade maior são vendidas por preços maiores, mantendo outros fatores iguais. Se o incinerador estivesse mais longe, em média, das casas melhores, a distância e a qualidade seriam positivamente correlacionadas, e portanto $\delta_1 > 0$. Uma regressão simples de *preço* sobre *distância* [ou log(*preço*) sobre log(*distância*)] tenderá a superestimar o efeito do incinerador: $\beta_1 + \beta_2\delta_1 > \beta_1$.

Suponha que o modelo

$$nota = \beta_0 + \beta_1 faltas + \beta_2 nmgradp + u$$

satisfaça as quatro primeiras hipóteses de Gauss-Markov, e onde *nota* é a nota de um exame final, *faltas* é o número de faltas e *nmgrad* é uma nota média acumulada até o penúltimo semestre. Se $\tilde{\beta}_1$ for o estimador de regressão simples de *nota* sobre *faltas*, qual será a direção do viés assimptótico em $\tilde{\beta}_1$?

Um ponto importante sobre a inconsistência dos estimadores de MQO é que, por definição, o problema não desaparece ao adicionarmos mais observações à amostra. O problema fica pior com mais dados: o estimador de MQO fica mais e mais próximo de $\beta_1 + \beta_2\delta_1$ quando aumenta o tamanho da amostra.

Derivar o sinal e a magnitude da inconsistência no caso geral de k regressores é mais difícil, do mesmo modo que derivar o viés é mais difícil. Precisamos lembrar que, se tivermos o modelo da equação (5.1), em que, por exemplo, x_1 é correlacionado com u, mas as outras variáveis independentes são não-correlacionadas com u, *todos* os estimadores de MQO serão geralmente inconsistentes. Por exemplo, no caso em que $k = 2$,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

suponha que x_2 e u sejam não-correlacionados, mas x_1 e u sejam correlacionados. Então, os estimadores de MQO $\hat{\beta}_1$ e $\hat{\beta}_2$ serão, em geral, inconsistentes. (O intercepto também será inconsistente.) A inconsistência em $\hat{\beta}_2$ surge quando x_1 e x_2 são correlacionados, como é normalmente o caso. Se x_1 e x_2 forem não-

correlacionados, então qualquer correlação entre x_1 e u *não* resulta na inconsistência de $\hat{\beta}_2$: plim $\hat{\beta}_2 = \beta_2$. Além disso, a inconsistência em $\hat{\beta}_1$ é a mesma que em (5.4). A mesma formulação se mantém no caso geral: se x_1 for correlacionado com u, mas x_1 e u não forem correlacionados com as outras variáveis independentes, então somente $\hat{\beta}_1$ é inconsistente, e a inconsistência é dada por (5.4). O caso geral é muito semelhante ao caso de variável omitida da Seção 3A.4 do Apêndice 3A, disponível no site da Thomson.

5.2 NORMALIDADE ASSIMPTÓTICA E INFERÊNCIA EM AMOSTRAS GRANDES

A consistência de um estimador é uma importante propriedade , mas ela sozinha não nos permite trabalhar com inferência estatística. Saber simplesmente que o estimador está se aproximando do valor populacional quando o tamanho da amostra cresce não nos permite testar hipóteses sobre os parâmetros. Para tanto, precisamos da distribuição amostral dos estimadores de MQO. Sob as hipóteses do modelo linear clássico, RLM.1 a RLM.6, o Teorema 4.1 mostra que as distribuições amostrais são normais. Esse resultado é a base para derivar as distribuições t e F que usamos com muita freqüência na econometria aplicada.

A normalidade exata dos estimadores de MQO depende crucialmente da normalidade da distribuição do erro, u, na população. Se os erros u_1, u_2, ..., u_n forem extrações aleatórias de alguma distribuição, diferente da normal, o $\hat{\beta}_j$ não será normalmente distribuído, o que significa que as estatísticas t não terão distribuições t e as estatísticas F não terão distribuições F. Esse é um problema potencialmente sério porque nossa inferência depende de sermos capazes de obter p-valores das distribuições t e F.

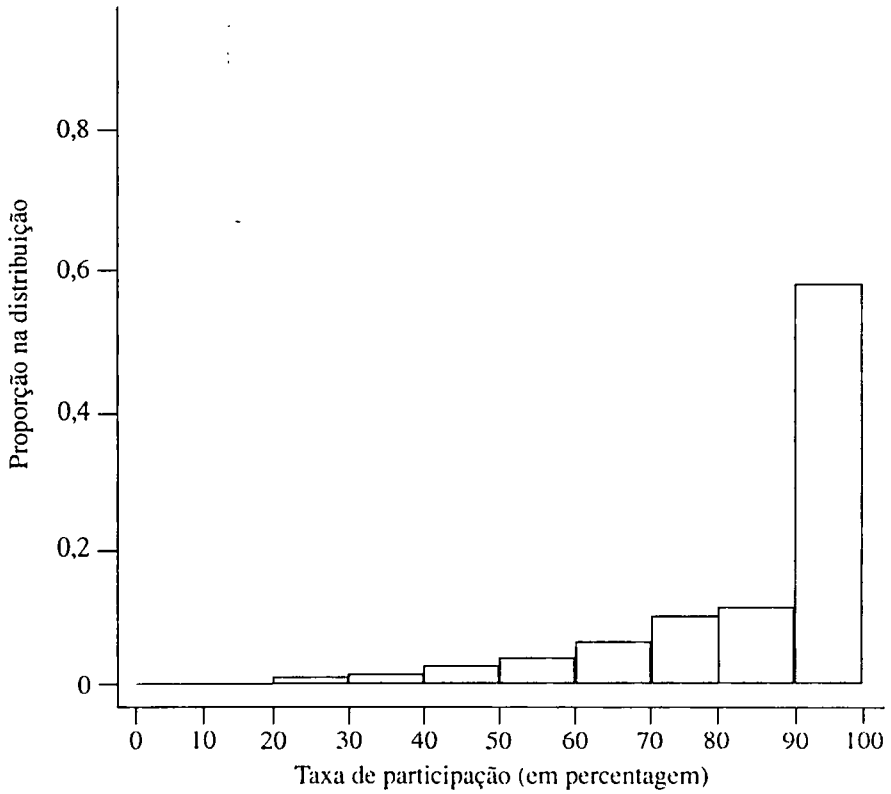
Lembre-se de que a hipótese RLM.6 é equivalente a dizer que a distribuição de y, dados x_1, x_2, ..., x_k, é normal. Como y é observado e u não é, em uma aplicação particular, é muito mais fácil pensar se é provável que a distribuição de y seja normal. De fato, já vimos alguns exemplos em que y definitivamente não poderia ter uma distribuição condicional normal. Uma variável aleatória normalmente distribuída é distribuída simetricamente ao redor de sua média, pode assumir qualquer valor positivo ou negativo (mas com probabilidade zero), e mais de 95% da área sob a distribuição está dentro de dois desvios-padrão.

No Exemplo 3.4 estimamos um modelo que explica o número de prisões de homens jovens durante um determinado ano (*npre86*). Na população, a maioria dos homens não estava presa durante o ano, e a maioria tinha sido presa uma vez no máximo. (Na amostra de 2.725 homens nos dados do arquivo CRIME1.RAW, menos de 8% foram presos mais que uma vez durante 1986.) Como *npre86* assume somente dois valores para 92% da amostra, ela não pode estar próxima de ser normalmente distribuída na população.

No Exemplo 4.6 estimamos um modelo que explica as percentagens de participação (*taxap*) nos planos de pensão nos Estados Unidos. A distribuição de freqüência (também chamada *histograma*) na Figura 5.2 mostra que a distribuição de *taxap* é fortemente inclinada para a direita, em vez de ser normalmente distribuída. De fato, mais de 40% das observações de *taxap* são iguais a 100, indicando participação de 100%. Isso viola a hipótese de normalidade, mesmo que condicional às variáveis explicativas.

Sabemos que a normalidade não tem nenhum papel na inexistência de viés de MQO, nem afeta a conclusão de que MQO é o melhor estimador linear não-viesado sob as hipóteses de Gauss-Markov. No entanto, a inferência exata baseada nas estatísticas t e F necessita de RML.6. Isso significa que, em nossa análise de *taxap* do Exemplo 4.6, devemos abandonar as estatísticas t para determinar quais variáveis são estatisticamente significantes? Felizmente, a resposta a essa questão é *não*. Ainda que os y_i não sejam provenientes de uma distribuição normal, podemos usar o teorema do limite central do Apêndice C, disponível no site da Thomson, para concluir que os estimadores de MQO satisfazem a **normalidade assimptótica**, o que significa que eles são, de maneira aproximada, normalmente distribuídos em amostras de tamanhos suficientemente grandes.

Figura 5.2
Histograma de *taxap*, usando dados do arquivo 401K.RAW.



TEOREMA 5.2 (NORMALIDADE ASSIMPTÓTICA DE MQO)
Sob as hipóteses de Gauss-Markov RLM.1 a RLM.5,

(i) $\sqrt{n}(\hat{\beta}_j - \beta_j) \overset{a}{\sim} \text{Normal}(0,\sigma^2/a_j^2)$, em que $\sigma^2/a_j^2 > 0$ é a **variância assimptótica** de $\sqrt{n}(\hat{\beta}_j - \beta_j)$; para os coeficientes de inclinação, $a_j^2 = \text{plim}\left(n^{-1}\sum_{i=1}^{n}\hat{r}_{ij}^2\right)$, em que os $\hat{r}_{ij}$ são os resíduos da regressão de x_j sobre outras variáveis independentes. Dizemos que $\hat{\beta}_j$ é assimptótica e normalmente distribuído (veja Apêndice C, disponível no site de Thomson);

(ii) $\hat{\sigma}^2$ é um estimador consistente de $\sigma^2 = \text{Var}(u)$;

(iii) Para cada j,

$$(\hat{\beta}_j - \beta_j)/\text{ep}(\hat{\beta}_j) \overset{a}{\sim} \text{Normal}(0,1), \qquad \textbf{(5.7)}$$

em que $\text{ep}(\hat{\beta}_j)$ é o erro-padrão usual de MQO.

A prova da normalidade assimptótica é um pouco complicada e está delineada para o caso da regressão linear simples no apêndice deste capítulo. A parte (ii) provém da lei dos grandes números e a parte (iii) decorre das partes (i) e (ii) e das propriedades assimptóticas discutidas no Apêndice C, disponível no site de Thomson.

O teorema 5.2 é útil porque a hipótese de normalidade RLM.6 foi excluída; a única restrição sobre a distribuição do erro é que ele tenha variância finita, algo que sempre assumiremos. Também assumimos média condicional zero e homoscedasticidade de u.

Observe como a distribuição normal padronizada aparece em (5.7), em oposição à distribuição t_{n-k-1}. Isso ocorre porque a distribuição é somente aproximada. Em contraste, no Teorema 4.2, a distribuição da relação em (5.7) era *exatamente* t_{n-k-1} para qualquer tamanho de amostra. De uma perspectiva prática, essa diferença é irrelevante. De fato, é válido escrever

$$(\hat{\beta}_j - \beta_j)/\text{ep}(\hat{\beta}_j) \overset{a}{\sim} t_{n-k-1}, \tag{5.8}$$

visto que t_{n-k-1} aproxima-se da distribuição normal padronizada quando os graus de liberdade tornam-se grandes.

A equação (5.8) nos diz que o teste t e a construção dos intervalos de confiança são realizados *exatamente* como sob as hipóteses do modelo linear clássico. Isso significa que nossa análise das variáveis dependentes, como *taxap* e *npre86*, não tem absolutamente de mudar se as hipóteses de Gauss-Markov se mantêm; em ambos os casos, temos pelo menos 1.500 observações, o que certamente é suficiente para justificar a aproximação pelo teorema do limite central.

Se o tamanho da amostra não é muito grande, então a distribuição t pode ser uma aproximação insatisfatória da distribuição da estatística t quando u não é normalmente distribuído. Infelizmente, antes de saber se a aproximação é suficientemente boa, não há prescrições gerais de quão grande deve ser o tamanho da amostra. Alguns econometristas pensam que $n = 30$ é satisfatório, mas esse valor pode não ser suficiente para todas as possíveis distribuições de u. Dependendo da distribuição de u, podem ser necessárias mais observações antes de o teorema do limite central começar a fazer efeito. Além disso, a qualidade da aproximação não depende apenas de n, mas dos gl, $n - k - 1$: com mais variáveis independentes no modelo, um tamanho de amostra maior é usualmente necessário para usar a aproximação t. Os métodos para inferência com graus de liberdade e erros não normais estão fora do escopo deste livro. Usaremos as estatísticas t como sempre usamos, sem nos preocuparmos com a hipótese de normalidade.

É muito importante ver que o Teorema 5.2, *de fato*, exige a hipótese de homoscedasticidade (juntamente com a hipótese de média condicional zero). Se $\text{Var}(y|\mathbf{x})$ não é constante, as estatísticas t usuais e os intervalos de confiança não são válidos, não importa quão grande seja o tamanho da amostra; na presença da heteroscedasticidade, o teorema do limite central em nada nos ajuda. Por essa razão, dedicaremos todo o Capítulo 8 à discussão do que pode ser feito na presença de heteroscedasticidade.

Uma conclusão do Teorema 5.2 é que $\hat{\sigma}^2$ é um estimador consistente de σ^2; já sabemos do Teorema 3.3 que $\hat{\sigma}^2$ é não-viesado para σ^2 sob as hipóteses de Gauss-Markov. A consistência implica que $\hat{\sigma}$ é um estimador consistente de σ, o que é importante para estabelecer o resultado da normalidade assimptótica na equação (5.7).

Lembre-se de que $\hat{\sigma}$ aparece no erro-padrão de cada $\hat{\beta}_j$. De fato, a variância estimada de $\hat{\beta}_j$ é

$$\hat{\text{Var}}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{\text{SQT}_j(1 - R_j^2)}, \tag{5.9}$$

em que SQT_j é a soma dos quadrados total de x_j na amostra, e R_j^2 é o R-quadrado da regressão de x_j sobre todas as outras variáveis independentes. Na Seção 3.4 estudamos cada componente de (5.9), os quais vamos expor agora no contexto da análise assimptótica. Quando o tamanho da amostra aumenta, $\hat{\sigma}^2$ converge em probabilidade para a constante σ^2. Além disso, R_j^2 se aproxima de um número estritamente entre zero e um (de modo que $1 - R_j^2$ converge para algum número entre zero e um). A variância amos-

tral de x_j é SQT_j/n, e portanto SQT_j/n converge para $Var(x_j)$ quando o tamanho da amostra aumenta. Isso quer dizer que SQT_j cresce aproximadamente à mesma taxa que o tamanho da amostra: $SQT_j \approx n\sigma_j^2$, em que σ_j^2 é a variância populacional de x_j. Quando combinamos esses fatos, vemos que $Vâr(\hat{\beta}_j)$ se contrai para zero à taxa de $1/n$; essa é a razão de tamanhos maiores de amostra serem melhores.

Em um modelo de regressão com um tamanho de amostra grande, qual é o intervalo de confiança de 95% aproximado para $\hat{\beta}_j$ sob RLM.1 a RLM.5? Ele é chamado de **intervalo de confiança assimptótico**.

Quando u não é normalmente distribuído, a raiz quadrada de (5.9) é, às vezes, chamada de **erro-padrão assimptótico**, e as estatísticas t são chamadas de **estatísticas t assimptóticas**. Como essas quantidades são as mesmas que estudamos no Capítulo 4, vamos chamá-las apenas de erros-padrão e estatísticas t, com o entendimento de que elas, algumas vezes, têm somente justificativa de amostra grande.

Ao usar o argumento anterior sobre a variância estimada, podemos escrever

$$ep(\hat{\beta}_j) \approx c_j / \sqrt{n}, \qquad \textbf{(5.10)}$$

em que c_j é uma constante positiva que *não* depende do tamanho da amostra. A equação (5.10) é somente uma aproximação, mas ela é uma regra de bolso útil: pode-se esperar que os erros-padrão diminuam a uma taxa que é o inverso da *raiz quadrada* do tamanho da amostra.

EXEMPLO 5.2

(Erros-Padrão em uma Equação do Peso de Nascimentos)

Usamos os dados do arquivo BWGHT.RAW para estimar uma relação em que o log do peso de nascimento é a variável dependente, e os cigarros fumados por dia (cigs) e o log da renda familiar são as variáveis independentes. O número total de observações é 1.388. Ao usar a primeira metade das observações (694), o erro-padrão de $\hat{\beta}_{cigs}$ é cerca de 0,0013. Ao usar todas as observações, o erro-padrão é cerca de 0,00086. A razão entre o último erro-padrão e o primeiro é 0,00086/0,0013 ≈ 0,662. Isso está bastante próximo de $\sqrt{694/1.388} \approx 0{,}707$, a razão obtida pela aproximação em (5.10). Em outras palavras, a equação (5.10) implica que o erro-padrão, ao usar o tamanho de amostra maior, deve ser 70,7% do erro-padrão obtido ao usar a amostra menor. Essa percentagem está muito perto dos 66,2% que realmente calculamos a partir da relação entre os erros-padrão.

A normalidade assimptótica dos estimadores de MQO também implica que as estatísticas F têm distribuições F aproximadas em tamanhos de amostras grandes. Assim, para testar as restrições de exclusão ou outras hipóteses múltiplas, nada muda em relação ao que tínhamos feito antes.

Outros Testes de Amostras Grandes: A Estatística Multiplicador de Lagrange

Visto que entramos no domínio da análise assimptótica, outras estatísticas de testes podem ser usadas para testar hipóteses. Para a maioria dos propósitos, há pouca razão para ir além das estatísticas t e F

usuais: como acabamos de ver, essas estatísticas têm justificativa de amostra grande sem a hipótese de normalidade. No entanto, algumas vezes é útil ter outras maneiras de testar restrições de exclusão múltiplas; por isso, vamos agora estudar a **estatística multiplicador de Lagrange (*LM*)**, que vem alcançando alguma popularidade na econometria moderna.

O nome "estatística multiplicador de Lagrange" provém da otimização com restrição, um tópico além do escopo deste livro. [Veja Davidson e MacKinnon (1993).] Também é usado o nome **estatística de escore** – o qual também é proveniente da otimização utilizada em cálculo. Felizmente, no arcabouço da regressão linear, é simples explicar a estatística *LM* sem se aprofundar na matemática mais complicada.

A forma da estatística *LM* que vamos derivar aqui apóia-se nas hipóteses de Gauss-Markov, as mesmas hipóteses que justificam a estatística *F* em amostras grandes. Não precisamos da hipótese de normalidade.

Para derivar a estatística *LM*, considere o modelo de regressão múltipla habitual com k variáveis independentes:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u. \qquad \textbf{(5.11)}$$

Gostaríamos de testar se, por exemplo, todas as últimas q dessas variáveis têm parâmetros populacionais zero: a hipótese nula é

$$H_0: \beta_{k-q+1} = 0, ..., \beta_k = 0, \qquad \textbf{(5.12)}$$

a qual coloca q restrições de exclusão sobre o modelo (5.11). Assim como no teste *F*, a hipótese alternativa a (5.12) é que pelo menos um dos parâmetros é diferente de zero.

A estatística *LM* exige somente a estimação do modelo *restrito*. Assim, assuma que computamos a regressão

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + ... + \tilde{\beta}_{k-q} x_{k-q} + \tilde{u}, \qquad \textbf{(5.13)}$$

em que "~" indica que as estimativas são do modelo restrito. Em particular, $\tilde{u}$ representa os resíduos do modelo restrito. (Como sempre, isso é apenas uma maneira de escrever para indicar que obtivemos o resíduo restrito de cada observação da amostra.)

Se as variáveis omitidas x_{k-q+1} até x_k tiverem, realmente, coeficientes populacionais zero, então $\tilde{u}$ deve ser, pelo menos aproximadamente, não-correlacionado com cada uma dessas variáveis na amostra. Isso sugere computar uma regressão desses resíduos sobre aquelas variáveis independentes excluídas sob H_0, que é semelhante ao que o teste *LM* faz. Entretanto, para obter uma estatística de teste que possa ser usada, devemos incluir *todas* as variáveis independentes na regressão. (A razão pela qual devemos incluir todos os regressores é que, em geral, os regressores omitidos no modelo restrito são correlacionados com os regressores que aparecem no modelo restrito.) Assim, computamos a regressão de

$$\tilde{u} \text{ sobre } x_1, x_2, ..., x_k. \qquad \textbf{(5.14)}$$

Esse é um exemplo de **regressão auxiliar**, uma regressão usada para calcular uma estatística de teste, mas cujos coeficientes não são de interesse direto.

Como podemos usar o resultado da regressão de (5.14) para testar (5.12)? Se (5.12) for verdadeira, o R-quadrado de (5.14) deve estar "próximo" de zero, sujeito ao erro amostral, porque $\tilde{u}$ será aproximadamente não-correlacionado com todas as variáveis independentes. A questão – como sempre ocorre com os testes de hipóteses – é como determinar quando a estatística é suficientemente grande para rejeitar a hipótese nula a um nível de significância escolhido. Isso resulta que, sob a hipótese nula, o tamanho da amostra multiplicado pelo R-quadrado da regressão auxiliar (5.14) é distribuído assimptoticamente como uma variável aleatória qui-quadrada com q graus de liberdade. Isso leva a um procedimento simples para testar a significância conjunta de um grupo de q variáveis independentes.

A ESTATÍSTICA MULTIPLICADOR DE LAGRANGE PARA q RESTRIÇÕES DE EXCLUSÃO:

(i) Regrida y sobre o conjunto *restrito* de variáveis independentes e salve os resíduos, $\tilde{u}$.
(ii) Regrida $\tilde{u}$ sobre *todas* as variáveis independentes e obtenha o R-quadrado, por exemplo R_u^2 (para distingui-lo dos R-quadrados obtidos com y como variável dependente).
(iii) Calcule $LM = nR_u^2$ [o tamanho amostral vezes o R-quadrado obtido no passo (ii)].
(iv) Compare o LM com o valor crítico apropriado, c, de uma distribuição χ_q^2; se $LM > c$, a hipótese nula é rejeitada. Melhor ainda, obtenha o p-valor como a probabilidade de que uma variável aleatória χ_q^2 exceda o valor da estatística de teste. Se o p-valor for menor que o nível de significância desejado, então H_0 é rejeitada. Se não for, não podemos rejeitar H_0. A regra de rejeição é essencialmente a mesma do teste F.

Por causa de sua forma, a estatística LM é às vezes referida como a **estatística *R*-*n*-quadrado**. Diferentemente da estatística F, os graus de liberdade do modelo irrestrito não têm qualquer papel na realização do teste LM. Tudo o que importa é o número de restrições que estão sendo testadas (q), o tamanho do R-quadrado auxiliar (R_u^2) e o tamanho da amostra (n). Os gl do modelo irrestrito não têm qualquer papel por causa da natureza assimptótica da estatística LM. No entanto, devemos estar certos de multiplicar R_u^2 pelo tamanho da amostra para obter LM; um valor aparentemente baixo do R-quadrado pode ainda levar à significância conjunta se n for grande.

Antes de dar um exemplo, uma palavra de precaução se faz necessária. Se, no passo (i), inadvertidamente regredirmos y sobre todas as variáveis independentes e utilizarmos os resíduos obtidos dessa regressão irrestrita no passo (ii), não vamos ter uma estatística interessante: o R-quadrado resultante será exatamente zero! Isso ocorre porque MQO escolhe as estimativas de modo que os resíduos sejam não-correlacionados nas amostras com todas as variáveis independentes incluídas [veja as equações (3.13)]. Assim, podemos somente testar (5.12) ao regredir os resíduos restritos sobre todas as variáveis independentes. (Regredir os resíduos restritos sobre o conjunto restrito de variáveis independentes também produzirá $R^2 = 0$.)

EXEMPLO 5.3

(Modelo Econômico do Crime)

Vamos ilustrar o teste LM ao usar uma versão ligeiramente mais extensa do modelo do crime do Exemplo 3.4:

$$npre86 = \beta_0 + \beta_1 pcond + \beta_2 sentmed + \beta_3 temptot + \beta_4 ptemp86 + \beta_5 empr86 + u,$$

em que *npre86* é o número de vezes que um homem foi preso, *pcond* é a proporção de prisões anteriores que levaram à condenação, *sentmed* é a sentença média cumprida de condenações passadas, *temptot* é o

EXEMPLO 5.3 (continuação)

tempo total que o homem passou na prisão em 1986 desde que atingiu a idade de 18 anos, *ptemp86* corresponde aos meses passados na prisão em 1986 e *empr86* é o número de trimestres, em 1986, durante os quais o homem esteve legalmente empregado. Vamos usar a estatística *LM* para testar a hipótese nula de que *sentmed* e *temptot* não têm efeito sobre *npre86*, uma vez que os outros fatores foram controlados.

No passo (i), estimamos o modelo restrito ao regredir *npre86* sobre *pcond*, *ptemp86* e *empr86* – as variáveis *sentmed* e *temptot* são excluídas dessa regressão – e obtemos os resíduos $\tilde{u}$ dessa regressão (2.725 resíduos). Então, computamos a regressão de

$$\tilde{u} \text{ sobre } pcond, ptemp86, empr86, sentmed \text{ e } temptot; \qquad \textbf{(5.15)}$$

como sempre, a ordem na qual listamos as variáveis independentes é irrelevante. Essa segunda regressão gera $R_{\tilde{u}}^2$, que é cerca de 0,0015. Esse valor pode parecer pequeno, mas devemos multiplicá-lo por n para obter a estatística $LM = 2.725(0,0015) \approx 4,09$. O valor crítico de 10% em uma distribuição qui-quadrada com dois graus de liberdade é cerca de 4,61 (arredondado para duas casas decimais; veja a Tabela G.4). Assim, não é possível rejeitar a hipótese nula de que $\beta_{sentmed} = 0$ e $\beta_{temptot} = 0$ ao nível de 10%. O *p*-valor é $P(\chi_2^2 > 4,09) \approx 0,129$, de modo que rejeitamos H_0 ao nível de 15%.

Como comparação, o teste *F* para a significância conjunta de *sentmed* e *temptot* resulta em um *p*-valor de cerca de 0,131, bastante próximo daquele obtido ao usar a estatística *LM*. Isso não é surpreendente, visto que assimptoticamente as duas estatísticas têm a mesma probabilidade de erro Tipo I. (Isto é, elas rejeitam a hipótese nula com a mesma freqüência quando a hipótese nula é verdadeira.)

Como o exemplo anterior sugere, com uma amostra grande raramente vemos discrepâncias importantes entre os resultados dos testes *LM* e *F*. Usaremos a estatística *F* para a maior parte dos problemas porque ela é rotineiramente calculada pela maioria dos pacotes de regressão. No entanto, você deve estar consciente sobre a estatística *LM* para reconhecê-la, pois ela é usada em trabalhos aplicados.

Um último comentário sobre a estatística *LM*. Assim como com a estatística *F*, devemos estar seguros de usar as mesmas observações nos passos (i) e (ii). Se faltarem dados para algumas das variáveis independentes excluídas sob a hipótese nula, os resíduos do passo (i) devem ser obtidos de uma regressão sobre o conjunto de dados reduzido.

5.3 EFICIÊNCIA ASSIMPTÓTICA DE MQO

Sabemos que, sob as hipóteses de Gauss-Markov, os estimadores de MQO são os melhores estimadores não-viesados lineares. MQO também é, sob as hipóteses de Gauss-Markov, **assimptoticamente eficiente** dentro uma classe de estimadores. Um tratamento geral requer álgebra matricial e análise assimptótica avançada. Em primeiro lugar, vamos descrever o resultado para o caso da regressão simples.

No modelo

$$y = \beta_0 + \beta_1 x + u, \qquad \textbf{(5.16)}$$

u tem média condicional zero sob RLM.3: $E(u|x) = 0$. Isso dá lugar a uma variedade de estimadores consistentes de β_0 e β_1; como habitual, vamos nos concentrar no parâmetro de inclinação, β_1. Seja $g(x)$ qual-

quer função de x; por exemplo, $g(x) = x^2$ ou $g(x) = 1/(1 + |x|)$. Então u é não-correlacionado com $g(x)$ (veja a Propriedade CE.5 no Apêndice B, disponível no site da Thomson). Seja $z_i = g(x_i)$ para todas as observações i. Então, o estimador

$$\tilde{\beta}_1 = \left(\sum_{i=1}^{n}(z_i - \bar{z})y_i\right)\Big/\left(\sum_{i=1}^{n}(z_i - \bar{z})x_i\right) \qquad \textbf{(5.17)}$$

é consistente para β_1, desde que $g(x)$ e x sejam correlacionados. [Lembre-se: é possível que $g(x)$ e x sejam não-correlacionados porque a correlação mensura a dependência *linear*.] Para ver isso, podemos colocar $y_i = \beta_0 + \beta_1 x_i + u_i$ em (5.17) e escrever $\tilde{\beta}_1$ como

$$\tilde{\beta}_1 = \beta_1 + \left(n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})u_i\right)\Big/\left(n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})x_i\right). \qquad \textbf{(5.18)}$$

Agora, podemos aplicar a lei dos grandes números ao numerador e denominador, os quais convergem em probabilidade para $\text{Cov}(z,u)$ e $\text{Cov}(z,x)$, respectivamente. Na condição de que $\text{Cov}(z,x) \neq 0$ – de modo que z e x sejam correlacionados –, temos

$$\text{plim}\,\tilde{\beta}_1 = \beta_1 + \text{Cov}(z,u)/\text{Cov}(z,x) = \beta_1,$$

porque $\text{Cov}(z,u) = 0$ sob RLM.3.

É mais difícil mostrar que $\tilde{\beta}_1$ é assimptoticamente normal. No entanto, usando argumentos semelhantes àqueles do apêndice deste capítulo, pode ser mostrado que $\sqrt{n}\,(\tilde{\beta}_1 - \beta_1)$ é assimptoticamente normal com média zero e variância assimptótica $\sigma^2\text{Var}(z)/[\text{Cov}(z,x)]^2$. A variância assimptótica do estimador de MQO é obtida quando $z = x$, caso em que $\text{Cov}(z,x) = \text{Cov}(x,x) = \text{Var}(x)$. Portanto, a variância assimptótica de $\sqrt{n}\,(\hat{\beta}_1 - \beta_1)$, em que $\hat{\beta}_1$ é o estimador de MQO, é $\sigma^2\text{Var}(x)/[\text{Var}(x)]^2 = \sigma^2/\text{Var}(x)$. Agora, a desigualdade de Cauchy-Schwartz (veja Apêndice B.4, disponível no site da Thomson) implica que $[\text{Cov}(z,x)]^2 \leq \text{Var}(z)\text{Var}(x)$, o que implica que a variância assimptótica de $\sqrt{n}\,(\hat{\beta}_1 - \beta_1)$ não é maior do que a de $\sqrt{n}\,(\tilde{\beta}_1 - \beta_1)$. Assim, para o caso da regressão simples, mostramos que, sob as hipóteses de Gauss-Markov, o estimador de MQO tem uma variância assimptótica menor do que qualquer outro estimador da forma (5.17). [O estimador em (5.17) exemplifica um *estimador de variáveis instrumentais*, que estudaremos extensivamente no Capítulo 15.] Se a hipótese de homoscedasticidade não for válida, então há estimadores da forma (5.17) que têm uma variância assimptótica menor do que a de MQO. Veremos isso no Capítulo 8.

O caso geral é semelhante, mas matematicamente muito mais difícil. No caso de k regressores, a classe de estimadores consistentes é obtida ao generalizar as condições de primeira ordem de MQO:

$$\sum_{i=1}^{n} g_j(\boldsymbol{x}_i)(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \ldots - \tilde{\beta}_k x_{ik}) = 0, j = 0, 1, \ldots, k, \qquad \textbf{(5.19)}$$

em que $g_j(\boldsymbol{x}_i)$ representa qualquer função de todas as variáveis explicativas para a observação i. Como pode ser visto ao comparar (5.19) com as condições de primeira ordem de MQO em (3.13), obtemos os estimadores de MQO quando $g_0(\boldsymbol{x}_i) = 1$ e $g_j(\boldsymbol{x}_i) = x_{ij}$, para $j = 1, 2, \ldots, k$. A classe dos estimadores em (5.19) é infinita, pois podemos usar qualquer função de x_{ij} que quisermos.

TEOREMA 5.3 (EFICIÊNCIA ASSIMPTÓTICA DE MQO)

Sob as hipóteses de Gauss-Markov, sejam $\tilde{\beta}_j$ os estimadores que solucionam as equações da forma (5.19) e sejam $\hat{\beta}_j$ os estimadores de MQO. Então, para j = 0, 1, 2, ..., *k*, os estimadores de MQO têm as menores variâncias assimptóticas: Avar $\sqrt{n}(\hat{\beta}_j - \beta_j) \leq$ Avar $\sqrt{n}(\tilde{\beta}_j - \beta_j)$.

Provar a consistência dos estimadores em (5.19), sem mostrar que eles são assimptoticamente normais, é matematicamente difícil. [Veja Wooldridge (2002, Capítulo 5).]

As afirmações subjacentes ao material deste capítulo são razoavelmente técnicas, mas suas implicações práticas são diretas. Mostramos que as quatro primeiras hipóteses de Gauss-Markov implicam que MQO é consistente. Além disso, todos os métodos de testar e construir intervalos de confiança que aprendemos no Capítulo 4 são aproximadamente válidos, sem assumir que os erros são extraídos de uma distribuição normal (equivalentemente, a distribuição de *y*, dadas as variáveis explicativas, não é normal). Isso significa que podemos aplicar MQO e usar os métodos anteriores para um conjunto de aplicações em que a variável dependente não é de fato aproximadamente normalmente distribuída. Também mostramos que, em vez da estatística *F*, a estatística *LM* pode ser usada para testar restrições de exclusão.

Antes de deixarmos este capítulo, devemos observar que coisas como o Exemplo 5.3 podem muito bem apresentar problemas que, *de fato*, exigem atenção especial. Para uma variável como *npre86*, que é zero ou um para a maioria dos homens na população, um modelo linear pode não ser capaz de adequadamente capturar a relação funcional entre *npre86* e as variáveis explicativas. Além do mais, mesmo se um modelo linear descreve o valor esperado das prisões, a heteroscedasticidade poderia ser um problema. Problemas como esses não são mitigados quando o tamanho da amostra aumenta, e portanto retornaremos a eles em capítulos posteriores.

5.1 No modelo de regressão simples sob RLM.1 a RLM.4, dissemos que o estimador de inclinação, $\hat{\beta}_1$, é consistente para β_1. Usando $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1$, mostre que plim $\hat{\beta}_0 = \beta_0$. [Você precisa usar a consistência de $\hat{\beta}_1$ e a lei dos grandes números, juntamente com o fato de que $\beta_0 = \mathrm{E}(y) - \beta_1\mathrm{E}(x_1)$.]

5.2 Suponha que o modelo

$$pctação = \beta_0 + \beta_1 funds + \beta_2 risctol + u$$

satisfaça as quatro primeiras hipóteses de Gauss-Markov, onde *pctação* é a percentagem da pensão de um trabalhador investida no mercado de ações, *funds* é o número de fundos mútuos que o trabalhador pode escolher e *risctol* é alguma medida de tolerância de risco (*risctol* maior significa que a pessoa tem uma tolerância maior ao risco). Se *funds* e *risctol* são positivamente correlacionados, qual é a inconsistência em $\tilde{\beta}_1$, o coeficiente de inclinação da regressão de *pctação* sobre *funds*?

5.3 O conjunto de dados do arquivo SMOKE.RAW contém informações sobre o comportamento tabagista e outras variáveis para uma amostra aleatória de adultos solteiros dos Estados Unidos. A variável *cigs* é o número (médio) de cigarros fumados por dia. Você acha que *cigs* tem uma distribuição normal na população adulta dos Estados Unidos? Explique.

5.4 No modelo de regressão simples (5.16), sob as quatro primeiras hipóteses de Gauss-Markov, mostramos que os estimadores da forma (5.17) são consistentes para a inclinação, β_1. Dado tal estimador, defina um estimador de β_0 como $\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x}$. Mostre que plim $\tilde{\beta}_0 = \beta_0$.

Vamos delinear uma prova da normalidade assimptótica de MQO [Teorema 5.2(i)] no caso da regressão simples. Escreva o modelo de regressão simples como na equação (5.16). Em seguida, por meio da álgebra usual da regressão simples, podemos escrever

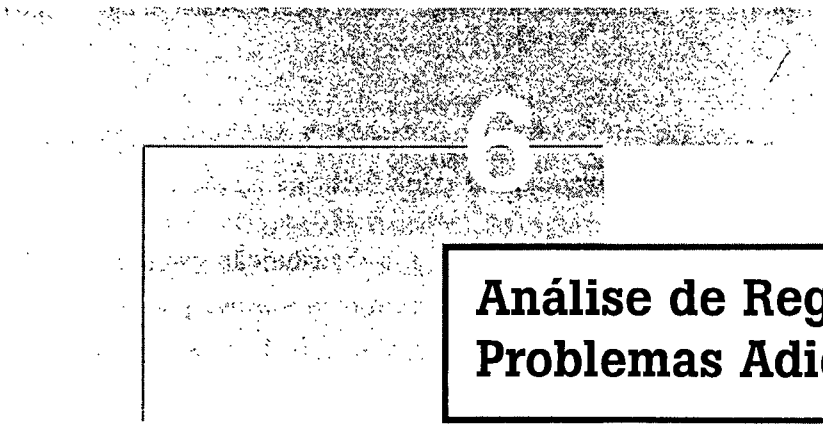
$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = (1/s_x^2)[n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i],$$

em que usamos s_x^2 para representar a variância amostral de $\{x_i: i = 1, 2, ..., n\}$. Pela lei dos grandes números (veja o Apêndice C, disponível no site de Thomson), $s_x^2 \xrightarrow{p} \sigma_x^2 = \text{Var}(x)$. A hipótese RLM.4 exclui a perfeita colinearidade, o que significa que $\text{Var}(x) > 0$ (x_i varia na amostra, e portanto x não é constante na população). Em seguida, $n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i = n^{-1/2}\sum_{i=1}^{n}(x_i - \mu)u_i + (\mu - \bar{x})[n^{-1/2}\sum_{i=1}^{n}u_i]$, em que $\mu = \text{E}(x)$ é a média populacional de x. Agora $\{u_i\}$ é a seqüência de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) com média zero e variância σ^2, e portanto $n^{-1/2}\sum_{i=1}^{n}u_i$ converge para a distribuição Normal(0, σ^2) quando $n \to \infty$; isso é exatamente o teorema do limite central do Apêndice C, disponível no site da Thomson. Pela lei dos grandes números, plim$(\mu - \bar{x}) = 0$. Um resultado padrão da teoria assimptótica é que se plim$(w_n) = 0$ e z_n tem uma distribuição normal assimptótica, então plim$(w_n z_n) = 0$. [Veja Wooldridge (2002, Capítulo 3) para mais discussão.] Isso implica que $(\mu - \bar{x})[n^{-1/2}\sum_{i=1}^{n}u_i]$ tem plim zero. Em seguida, $\{(x_i - \mu)u_i: i = 1, 2, ...\}$ é uma seqüência indefinida de variáveis aleatórias i.i.d. com média zero – porque u e x são não-correlacionados sob RLM.3 – e variância $\sigma^2\sigma_x^2$, pela hipótese de homoscedasticidade RLM.5. Portanto, $n^{-1/2}\sum_{i=1}^{n}(x_i - \mu)u_i$ tem uma distribuição Normal(0, $\sigma^2\sigma_x^2$) assimptótica. Acabamos de mostrar que a diferença entre $n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i$ e $n^{-1/2}\sum_{i=1}^{n}(x_i - \mu)u_i$ tem plim zero. Um resultado da teoria assimptótica é que se z_n tem uma distribuição normal e plim$(v_n - z_n) = 0$, então v_n tem a mesma distribuição normal assimptótica que z_n. Em decorrência $n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i$ também tem uma distribuição Normal(0, $\sigma^2\sigma_x^2$) assimptótica. Colocando todas essas peças juntas, temos

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = (1/\sigma_x^2)[n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i]$$

$$+ [(1/s_x^2) - (1/\sigma_x^2)][n^{-1/2}\sum_{i=1}^{n}(x_i - \bar{x})u_i],$$

e como plim$(1/s_x^2) = 1/\sigma_x^2$, o segundo termo tem plim zero. Portanto, a distribuição assimptótica de $\sqrt{n}(\hat{\beta}_1 - \beta_1)$ é Normal(0, $\{\sigma^2\sigma_x^2\}/\{\sigma_x^2\}^2$) = Normal(0, σ^2/σ_x^2). Isso completa a prova para o caso da regressão simples, quando $a_1^2 = \sigma_x^2$ neste caso. Veja Wooldridge (2002, Capítulo 4) para o caso geral.



Análise de Regressão Múltipla: Problemas Adicionais

ste capítulo articula vários problemas da análise de regressão múltipla que não foram tratados convenientemente nos capítulos anteriores. Estes tópicos não são tão fundamentais quanto os discutidos nos capítulos 3 e 4, mas são importantes para a aplicação da regressão múltipla em uma ampla gama de problemas empíricos.

6.1 EFEITOS DA DIMENSÃO DOS DADOS NAS ESTATÍSTICAS MQO

No Capítulo 2, sobre regressão bivariada, discutimos de forma sucinta os efeitos da mudança nas unidades de medida sobre os interceptos e as estimativas de inclinação do MQO. Também mostramos que a mudança nas unidades de medida não afeta o R-quadrado. Agora retornaremos ao problema da dimensão dos dados e examinaremos o efeito do redimensionamento das variáveis dependente ou independente sobre os erros-padrão, estatísticas t, estatísticas F e intervalos de confiança.

Veremos que tudo o que esperamos acontecer, acontecerá. Quando as variáveis são redimensionadas, os coeficientes, erros-padrão, intervalos de confiança, estatísticas t e F mudam de tal maneira que preservam todos os efeitos mensurados e os resultados dos testes. Embora isso não seja uma grande surpresa – aliás, ficaríamos muito preocupados se não fosse assim – é útil ver o que ocorre explicitamente. Muitas vezes, o redimensionamento de dados é usado com finalidade cosmética, tal como reduzir o número de zeros depois da vírgula, em um coeficiente estimado. Escolhendo-se criteriosamente as unidades de medida, podemos melhorar a aparência de uma equação estimada sem alterar nada que seja essencial.

Poderíamos tratar deste problema de maneira generalizada, mas ele é mais bem ilustrado com exemplos. Da mesma forma, será de pouca valia neste ponto introduzirmos uma notação abstrata.

Começamos com uma equação relacionando o peso dos recém-nascidos com o hábito de fumar e a renda familiar:

$$\widehat{pesônas} = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 rendfam \qquad \textbf{(6.1)}$$

onde *pesônas* é o peso dos recém-nascidos, em onças, *cigs* é o número médio de cigarros que a mãe fumou por dia durante a gravidez, e *rendfam* é a renda anual familiar, em milhares de dólares. As estimativas desta equação, obtidas utilizando dados contidos no arquivo BWGHT.RAW, são dadas na primeira coluna da Tabela 6.1. Os erros-padrão estão relacionados entre parênteses. A estimativa de *cigs* mostra que se uma mulher fumar cinco ou mais cigarros por dia, o peso previsto dos recém-

nascidos deve estar em torno de 0,4634(5) = 2,317 onças a menos. A estatística *t* de *cigs* é −5,06, de modo que a variável é estatisticamente bastante significante.

Tabela 6.1

Efeitos da Dimensão dos Dados

Variável Dependente	(1) *pesonas*	(2) *pesonaslb*	(3) *pesonas*
Variáveis Independentes			
cigs	−0,4634 (0,0916)	−0,0289 (0,0057)	———
maços	———	———	−9,268 (1,832)
rendfam	0,0927 (0,0292)	0,0058 (0,0018)	0,0927 (0,0292)
intercepto	116,974 (1,049)	7,3109 (0,0656)	116,974 (1,049)
Observações	1.388	1.388	1.388
R-quadrado	0,0298	0,0298	0,0298
SQR	557.485,51	2.177,6778	557.485,51
EPR	20,063	1,2539	20,063

Agora, suponha que decidimos medir o peso dos recém-nascidos em libras, em vez de onças. Façamos *pesonaslb* = *pesonas*/16 ser o peso dos recém-nascidos em libras. O que acontece com nossas estatísticas MQO se usarmos essa variável dependente em nossa equação? É fácil verificar o efeito no coeficiente da estimativa pela simples manipulação da equação (6.1). Divida a equação inteira por 16:

$$pesônas/16 = \hat{\beta}_0/16 + (\hat{\beta}_1/16)cigs + (\hat{\beta}_2/16)rendfam.$$

Como o termo da esquerda é o peso dos recém-nascidos em libras, segue-se que cada novo coeficiente corresponderá ao coeficiente antigo dividido por 16. Para verificar isso, a regressão de *pesonaslb* sobre *cigs* e *rendfam* está registrada na coluna (2) da Tabela 6.1. Até quatro dígitos, o intercepto e as inclinações da coluna (2) são exatamente os da coluna (1) divididos por 16. Por exemplo, o coeficiente de *cigs* é agora −0,0289; isso significa que, se *cigs* fosse cinco vezes mais alto, o peso de nascimento seria 0,0289(5) = 0,1445 *libras* mais baixo. Em termos de onças, temos 0,1445(16) = 2,312, que é um pouco diferente dos 2,317 que obtivemos anteriormente devido ao erro de arredondamento. A questão importante é que, uma vez que os efeitos tenham sido transformados nas mesmas unidades, obtemos exatamente a mesma resposta, independentemente de como a variável dependente seja medida.

E quanto à significância estatística? Como esperado, a alteração da variável dependente de onças para libras não tem efeito sobre o quanto são estatisticamente importantes as variáveis independentes. Os erros-padrão na coluna (2) são 16 vezes menores que os da coluna (1). Alguns cálculos rápidos mostram que as estatísticas t na coluna (2) são, realmente, idênticas às da coluna (1). Os pontos extremos dos intervalos de confiança na coluna (2) são exatamente os pontos extremos na coluna (1) divididos por 16. Isso ocorre porque os ICs mudam pelos mesmos fatores dos erros-padrão. (Lembre-se de que o IC de 95% neste caso é $\hat{\beta}_j \pm 1{,}96\ \text{ep}(\hat{\beta}_j)$.

Em termos de grau de ajuste, os R-quadrados das duas regressões são idênticos, como esperado. Observe que a soma dos resíduos quadrados, SQR, e o erro-padrão da regressão, EPR, diferem nas equações. Essas diferenças são facilmente explicadas. Seja $\hat{u}_i$ o resíduo da observação i na equação original (6.1). Então, quando *pesonaslbs* é a variável dependente, o resíduo é simplesmente $\hat{u}_i/16$. Assim, o resíduo *quadrado* na segunda equação é $(\hat{u}_i/16)^2 = \hat{u}_i^2/256$. Essa é a razão pela qual a soma dos resíduos quadrados na coluna (2) é igual à SQR na coluna (1) dividida por 256.

Como EPR $= \hat{\sigma} = \sqrt{\text{SQR}/(n-k-1)} = \sqrt{\text{SQR}/1.385}$, SQR na coluna (2) é 16 vezes menor do que na coluna (1). Outra maneira de ver isso é que o erro na equação com *pesonaslb* como a variável dependente tem um desvio-padrão 16 vezes menor do que o desvio-padrão do erro original. Isso não significa que tenhamos reduzido o erro ao alterarmos a maneira pela qual o peso dos recém-nascidos é medido: o EPR menor simplesmente reflete uma diferença nas unidades de medida.

Continuando, retornemos à unidade de medida original da variável dependente: *pesonas* é medido em onças. Vamos alterar a unidade de medida de uma das variáveis independentes, *cigs*. Defina *maços* como sendo a quantidade de maços de cigarros fumados por dia. Assim, *maços* = *cigs*/20. Agora, o que acontece com os coeficientes e outras estatísticas MQO? Dessa forma, podemos escrever

$$pesônas = \hat{\beta}_0 + (20\hat{\beta}_1)(cigs/20) + \hat{\beta}_2 rendfam = \hat{\beta}_0 + (20\hat{\beta}_1)maços + \hat{\beta}_2 rendfam.$$

Portanto, o intercepto e o coeficiente de inclinação de *rendfam* não se alteraram, mas o coeficiente de *maços* é 20 vezes o de *cigs*. Isso é intuitivamente atraente. Os resultados da regressão de *pesonas* sobre *maços* e *rendfam* estão na coluna (3) da Tabela 6.1. A propósito, lembre-se de que não teria sentido incluir tanto *cigs* como *maços* na mesma equação; isso induziria à multicolinearidade perfeita e não teria nenhum significado interessante.

Na equação original sobre o peso dos recém-nascidos (6.1), suponha que *rendfam* seja medida em dólares em lugar de milhares de dólares. Desse modo, defina a variável *rendfamdol* = 1.000 · *rendfam*. Como mudam as estatísticas MQO quando *rendfamdol* substitui *rendfam*? Para o propósito de apresentar os resultados da regressão, você acha melhor medir a renda em dólares ou em milhares de dólares?

Além do coeficiente de *maços*, existe outra estatística na coluna (3) que difere da mostrada na coluna (1): o erro-padrão de *maços* é 20 vezes maior que o de *cigs* na coluna (1). Isso significa que a estatística t para verificar a significância do hábito de fumar é a mesma, quer ele seja medido em cigarros ou em maços. Isso é natural.

O exemplo anterior explica claramente a maioria das possibilidades que surgem quando a variável dependente e as variáveis independentes são redimensionadas. O redimensionamento muitas

vezes é feito com os valores monetários em economia, especialmente quando os montantes são muito grandes.

No Capítulo 2, argumentamos que, se a variável dependente aparecer na forma logarítmica, a alteração na unidade de medida não afetará o coeficiente de inclinação. Isso também acontece aqui: a alteração na unidade de medida da variável dependente, quando aparece na forma logarítmica, não afeta qualquer das estimativas de inclinação. Isso resulta do simples fato de que $\log(c_1,y_i) = \log(c_1) + \log(y_i)$ para qualquer constante $c_1 > 0$. O novo intercepto será $\log(c_1) + \hat{\beta}_0$. De forma semelhante, a alteração da unidade de medida de qualquer x_j, onde $\log(x_j)$ aparece na regressão, afeta somente o intercepto. Isso corresponde ao que conhecemos sobre alterações em porcentagens e, em particular, em elasticidades: elas não sofrem alterações quando mudam as unidades de medida de y ou de x_j. Por exemplo, se tivéssemos especificado a variável dependente em (6.1) como log(*pesonas*), estimássemos a equação, e depois a tivéssemos reestimado com log(*pesonaslb*) como a variável dependente, os coeficientes de *cigs* e *rendfam* seriam os mesmos em ambas as regressões; somente o intercepto seria diferente.

Os Coeficientes Beta

Algumas vezes, em aplicações econométricas, uma variável-chave é medida em uma dimensão de difícil interpretação. Economistas especializados na área de trabalho freqüentemente incluem a pontuação de testes de conhecimentos em equações salariais, e a dimensão em que tais testes são registrados muitas vezes é arbitrária e de difícil interpretação (pelo menos para os economistas!). Em quase todos os casos estamos interessados em saber como a pontuação de um indivíduo em particular se compara com a população. Assim, em lugar de perguntarmos a respeito do efeito sobre o salário por hora se, digamos, a pontuação do teste for dez pontos mais alta, faz mais sentido perguntar o que acontece quando a pontuação do teste for um desvio-padrão mais alto.

Nada impede que vejamos o que acontece com a variável dependente quando uma variável independente em um modelo estimado aumenta certo número de desvios-padrão, supondo que tenhamos obtido o desvio-padrão da amostra (o que é fácil na maioria dos programas de regressão). Geralmente, essa é uma boa idéia. Assim, por exemplo, quando observamos o efeito de uma pontuação de teste padronizada, como o SAT (nota de ingresso em curso superior nos Estados Unidos), sobre a nota média em curso superior, podemos encontrar o desvio-padrão de SAT e verificar o que acontece quando essa pontuação aumenta em um ou dois desvios-padrão.

Algumas vezes é útil obter resultados de regressão quando todas as variáveis envolvidas, a dependente e todas as independentes, tenham sido padronizadas. Uma variável é padronizada em uma amostra pela subtração de sua média e dividindo o resultado por seu desvio-padrão (veja Apêndice C disponível no site do livro, no site www.thomsonlearning.com.br). Isso significa que computamos a transformação z de cada variável na amostra. Depois, fazemos a regressão usando os valores de z.

Por que a padronização é útil? É mais fácil começarmos com a equação MQO original, com as variáveis em suas formas originais:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}. \qquad \textbf{(6.2)}$$

Incluímos o subscrito de observação i para enfatizar que nossa padronização é aplicada a todos os valores da amostra. Agora, se ao calcularmos a média de (6.2), usarmos o fato de que $\hat{u}_i$ tem uma média de amostra zero, e subtrairmos o resultado de (6.2), temos

$$y_i - \bar{y} = \hat{\beta}_1(x_{i1} - \bar{x}_1) + \hat{\beta}_2(x_{i2} - \bar{x}_2) + \ldots + \hat{\beta}_k(x_{ik} - \bar{x}_k) + \hat{u}_i.$$

Em seguida, definamos $\hat{\sigma}_y$ como o desvio-padrão da amostra da variável dependente, $\hat{\sigma}_1$ como o *dp* da amostra da x_1, $\hat{\sigma}_2$ como o *dp* da amostra de x_2, e assim sucessivamente. Agora, um pouco de álgebra produz a equação

$$(y_i - \bar{y})/\hat{\sigma}_y = (\hat{\sigma}_1/\hat{\sigma}_y)\hat{\beta}_1[(x_{i1} - \bar{x}_1)/\hat{\sigma}_1] + ...$$
$$+ (\hat{\sigma}_k/\hat{\sigma}_y)\hat{\beta}_k[(x_{ik} - \bar{x}_k)/\hat{\sigma}_k] + (\hat{u}_i/\hat{\sigma}_y). \quad \textbf{(6.3)}$$

Cada variável em (6.3) foi padronizada pela substituição de suas médias por seus valores de *z*, e isso resultou em novos coeficientes de inclinação. Por exemplo, o coeficiente de inclinação de $(x_{i1} - \bar{x}_1)/\hat{\sigma}_1$ é $(\hat{\sigma}_1/\hat{\sigma}_y)\hat{\beta}_1$. Isso é simplesmente o coeficiente original, $\hat{\beta}_1$, multiplicado pela razão do desvio-padrão de x_1 sobre o desvio-padrão de *y*. O intercepto simplesmente desapareceu.

É útil reescrever (6.3), eliminando o subscrito *i*, como

$$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 + ... + \hat{b}_k z_k + erro, \quad \textbf{(6.4)}$$

onde z_y é o valor de *z* de *y*, z_1 é o valor de *z* de x_1, e assim por diante. Os novos coeficientes são

$$\hat{b}_j = (\hat{\sigma}_j/\hat{\sigma}_y)\hat{\beta}_j \text{ para } j = 1, ..., k. \quad \textbf{(6.5)}$$

Esses $\hat{b}_j$ são tradicionalmente chamados de **coeficientes padronizados** ou **coeficientes beta**. (Esta última denominação é mais comum, mas um pouco inadequada, já que temos usado o beta chapéu para representar as estimativas MQO *usuais*.)

Os coeficientes beta recebem seus interessantes significados a partir da equação (6.4): Se x_1 aumentar em um desvio-padrão, $\hat{y}$, então, será alterado em $\hat{b}_1$ desvios-padrão. Assim, estamos medindo os efeitos não em termos das unidades originais de *y* ou de x_j, mas em unidades de desvios-padrão. Como isso torna a dimensão dos regressores irrelevante, essa equação coloca as variáveis explicativas em pé de igualdade. Em uma equação MQO padrão, não é possível simplesmente verificar o tamanho dos diferentes coeficientes e concluir que a variável explicativa com o maior coeficiente é "a mais importante". Acabamos de ver que a magnitude dos coeficientes pode ser mudada à vontade pela alteração das unidades de medida das variáveis x_j. Mas, quando cada x_j é padronizado, a comparação das magnitudes dos coeficientes beta resultantes é mais convincente.

Para obter os coeficientes beta, podemos sempre padronizar *y*, x_1,..., x_k e em seguida computar a regressão MQO do valor de *z* de *y* sobre os valores de *z* de x_1,..., x_k – no qual não é necessário incluir um intercepto, já que ele será zero. Isso pode ser tedioso com muitas variáveis independentes. Alguns programas econométricos produzem coeficientes beta com um simples comando. O exemplo seguinte ilustra o uso de coeficientes beta.

EXEMPLO 6.1

(Efeitos da Poluição sobre os Preços de Imóveis)

Utilizamos os dados do Exemplo 4.5 (do arquivo HPRICE2.RAW) para ilustrar o uso de coeficientes beta. Lembre-se de que a principal variável independente é *oxn*, uma medida do óxido nitroso no ar em cada comunidade. Uma maneira de entender o tamanho do efeito da poluição – sem entrar na questão científica do efeito do óxido de nitrogênio sobre a qualidade do ar é computar os coeficientes beta. (O Exemplo 4.5 contém um método alternativo: obtivemos uma elasticidade-preço em relação a *oxn* usando *preço* e *oxn* em forma logarítmica.)

A equação populacional é o modelo nível-nível

$$preço = \beta_0 + \beta_1 oxn + \beta_2 crime + \beta_3 comods + \beta_4 dist + \beta_5 razestud + u,$$

onde todas as variáveis exceto *crime* foram definidas no Exemplo 4.5; *crime* é o número de crimes registrados *per capita*. Os coeficientes beta aparecem na seguinte equação (portanto cada variável foi convertida ao seu valor de *z*):

$$\widehat{zpreço} = -0{,}340\ zoxn - 0{,}143\ zcrime + 0{,}514\ zcomods - 0{,}235\ zdist - 0{,}270\ zrazestud.$$

Esta equação mostra que o aumento de um desvio-padrão em *oxn* reduz o preço em 0,34 desvio-padrão; o aumento de um desvio-padrão em *crime* reduz o preço em 0,14 desvio-padrão. Assim, o mesmo movimento relativo da poluição na população tem um efeito maior sobre os preços dos imóveis do que o da criminalidade. O tamanho do imóvel, medido pelo número de cômodos (*comods*), tem o maior efeito padronizado. Se quisermos saber os efeitos de cada variável independente sobre o valor da mediana dos preços dos imóveis, teremos que usar as variáveis não padronizadas.

O uso de variáveis padronizadas ou não padronizadas não afetará a significância estatística: as estatísticas *t* serão as mesmas, em ambos os casos.

6.2 UM POUCO MAIS SOBRE A FORMA FUNCIONAL

Em vários dos exemplos anteriores, encontramos o artifício mais comum em econometria para permitir relações não lineares entre a variável explicada e as variáveis explicativas: o uso de logaritmos das variáveis dependentes ou independentes. Também vimos modelos contendo os quadrados de algumas variáveis explicativas, mas ainda precisamos discorrer sobre um tratamento sistemático desses tópicos. Nesta seção, trataremos de algumas variações e extensões sobre formas funcionais que surgem freqüentemente em trabalhos aplicados.

Um pouco mais sobre o Uso de Formas Funcionais Logarítmicas

Começamos revendo como interpretar os parâmetros no modelo

$$\log(preço) = \beta_0 + \beta_1 \log(oxn) + \beta_2 comods + u, \qquad \textbf{(6.6)}$$

onde essas variáveis são as mesmas do Exemplo 4.5. Lembre-se de que em todo o texto log(x) é o log *natural* de x. O coeficiente β_1 é a elasticidade do *preço* em relação a *oxn* (poluição). O coeficiente β_2 é a mudança em log(*preço*), quando $\Delta comods = 1$; como vimos muitas vezes, quando multiplicada por 100, essa é a percentagem aproximada de mudança em *preço*. Lembre-se de que $100 \cdot \beta_2$ é algumas vezes chamado de semi-elasticidade do *preço* em relação a *comods*.

Quando estimamos utilizando os dados do arquivo HPRICE2.RAW, obtemos

$$\widehat{\log(preço)} = 9{,}23 - 0{,}718 \log(oxn) + 0{,}306\, comods \quad \textbf{(6.7)}$$
$$(0{,}19)\ \ (0{,}66)\qquad\qquad (0{,}019)$$
$$n = 506,\ R_2 = 0{,}514.$$

Assim, quando *oxn* aumenta em 1%, *preço* cai em 0,718%, mantendo-se apenas *comods* fixo. Quando *comods* aumenta em um, *preço* aumenta em aproximadamente 100(0,306) = 30,6%.

A estimativa de que um cômodo a mais aumenta o preço em cerca de 30,6% acaba por ser de certa forma impreciso para esta aplicação. O erro de aproximação ocorre porque, como a mudança em log(y) se torna cada vez maior, a aproximação $\%\Delta y \approx 100 \cdot \Delta\log(y)$ se mostra cada vez mais imprecisa. Felizmente, existe um cálculo simples para computar a percentagem exata de mudança.

Para descrever o procedimento, consideremos o modelo estimado de forma geral

$$\hat{\log}(y) = \hat{\beta}_0 + \hat{\beta}_1 \log(x_1) + \hat{\beta}_2 x_2.$$

(A inclusão de variáveis independentes adicionais não altera o procedimento.) Agora, fixando x_1, temos $\Delta\hat{\log}(y) = \hat{\beta}_2 \Delta x_2$. O uso de simples propriedades algébricas das funções exponenciais e logarítmicas produz a percentagem exata de mudança no y estimado como

$$\%\,\hat{\Delta} y = 100 \cdot [\exp(\hat{\beta}_2 \Delta x_2) - 1], \quad \textbf{(6.8)}$$

onde a multiplicação por 100 transforma a mudança proporcional em uma mudança percentual. Quando $\Delta x_2 = 1$,

$$\%\,\hat{\Delta} y = 100 \cdot [\exp(\hat{\beta}_2) - 1]. \quad \textbf{(6.9)}$$

Aplicada ao exemplo dos preços dos imóveis com $x_2 = comods$ e $\hat{\beta}_2 = 0{,}306$, $\%\Delta\widehat{preço} = 100[\exp(0{,}306) - 1] = 35{,}8\%$, que é visivelmente maior do que a percentagem aproximada de mudança, 30,6%, obtida diretamente de (6.7). {A propósito, esse não é um estimador não-viesado, pois exp(·) é uma função não-linear; ele é, porém, um estimador consistente de $100[\exp(\beta_2) - 1]$. Isso é assim porque o limite de probabilidade é calculado por meio de funções contínuas, enquanto o operador valor esperado não é calculado dessa forma. Veja Apêndice C, no site da Thomson.}

O ajuste na equação (6.8) não é tão crucial para pequenas mudanças percentuais. Por exemplo, quando incluímos na equação (6.7) a relação aluno-professor, seu coeficiente estimado é −0,052, o que significa que, se *razestud* aumentar em um, *preço* diminui em aproximadamente 5,2%. A mudança proporcional exata é $\exp(-0{,}052) - 1 \approx -0{,}051$, ou −5,1%. De outro lado, se aumentarmos *razestud* em

cinco, então a mudança percentual aproximada em *preço* será −26%, enquanto a mudança exata obtida da equação (6.8) é $100[\exp(-0,26) - 1] \approx -22,9\%$.

Vimos que o uso de logs naturais leva a coeficientes com interpretações interessantes e podemos ignorar o fato de as unidades de medida das variáveis aparecerem em forma logarítmica, pois os coeficientes de inclinação são invariantes em relação a redimensionamentos. Existem várias outras razões pelas quais os logs são tão usados em trabalhos aplicados. Em primeiro lugar, quando $y > 0$, os modelos que usam $\log(y)$ como a variável dependente geralmente satisfazem as hipóteses do MLC mais apropriadamente do que os modelos que usam o nível de y. Variáveis estritamente positivas freqüentemente possuem distribuições condicionais que são heteroscedásticas ou concentradas; o uso do log pode aliviar, se não eliminar, ambos os problemas.

Além disso, o uso de logs normalmente estreita a amplitude dos valores das variáveis, em alguns casos em quantidade considerável. Isso torna as estimativas menos sensíveis a observações díspares (ou extremas) na variável dependente ou nas variáveis independentes. Abordaremos a questão das observações extremas no Capítulo 9.

Existem algumas regras práticas padronizadas para o uso de logs, embora nenhuma definitiva. Quando a variável é um valor monetário positivo, ele freqüentemente é transformado em log. Temos visto isso para variáveis como salários, vendas de empresas e valores de mercado das empresas. Variáveis como população, número total de empregados e matrículas escolares freqüentemente aparecem em forma logarítmica; elas têm a característica comum de serem grandes valores inteiros.

Variáveis que são medidas em anos — como educação, experiência, tempo de permanência, idade etc. — normalmente aparecem em sua forma original. Uma variável que seja uma proporção ou uma percentagem — como a taxa de desemprego, a taxa de participação em planos de aposentadoria, a taxa de estudantes aprovados em um exame padronizado e a taxa de detenção sobre crimes registrados — pode aparecer tanto em sua forma original como logarítmica, embora haja uma tendência em usá-la em forma de nível. Isso se deve ao fato de que quaisquer coeficientes de regressão envolvendo a variável *original* — seja ela a variável dependente ou independente — terão uma interpretação de mudança de *pontos percentuais*. (Veja Apêndice A, no site da Thomson, para uma revisão sobre a distinção entre mudança percentual e mudança de pontos percentuais.) Se usarmos, digamos, $\log(desemp)$ em uma regressão, onde *desemp* é a percentagem de indivíduos desempregados, precisamos ter muito cuidado para distinguir entre uma mudança de pontos percentuais e uma mudança percentual. Lembre-se, quando *desemp* aumenta de oito para nove, isso é um acréscimo de um ponto percentual, equivalente a um incremento de 12,5% sobre o nível de desemprego inicial. O uso do log significa que queremos saber a mudança percentual da taxa de desemprego: $\log(9) - \log(8) \approx 0,118$ ou 11,8%, que é a aproximação logarítmica do aumento efetivo de 12,5%.

Suponha que o número anual de prisões por direção de veículo sob embriaguez[1] seja determinado por

$$\log(prisões) = \beta_0 + \beta_1 \log(pop) + \beta_2 idade16\_25 + outros\ fatores,$$

onde *idade16_25* é a proporção da população entre 16 e 25 anos de idade. Mostre que β_2 tem a seguinte interpretação (*ceteris paribus*): ela é a mudança percentual em *prisões* quando a percentagem da população com idade entre 16 e 25 anos aumenta em um *ponto percentual*.

[1] NE: Nos Estados Unidos.

Uma limitação do log é que ele não pode ser usado, caso uma variável assuma valor zero ou negativo. Em casos nos quais a variável y não seja negativa, mas pode assumir o valor 0, $\log(1 + y)$ é algumas vezes usado. As interpretações de mudança percentual são, em geral, estritamente preservadas, exceto para mudanças começando em $y = 0$ (em que a percentagem de mudança não é sequer definida). Geralmente, usar $\log(1 + y)$ e depois interpretar as estimativas como se a variável fosse $\log(y)$ é aceitável quando os dados em y não são dominados por zeros. Um exemplo pode ser o de y representar horas de treinamento por funcionário da população industrial, quando uma grande fração das empresas oferece treinamento a, pelo menos, um de seus funcionários.

Uma desvantagem de usar uma variável dependente na forma logarítmica está na dificuldade de se prever a variável original. O modelo original nos permite prever $\log(y)$, e não y. No entanto, é razoavelmente fácil transformar uma previsão de $\log(y)$ em uma previsão de y (veja Seção 6.4). Uma questão relacionada é que *não* é válido comparar R-quadrados de modelos nos quais y é a variável dependente em um caso e $\log(y)$ é a variável dependente no outro. Essas medidas explicam variações em diferentes variáveis. Discutimos como computar medidas comparáveis de graus de ajuste na Seção 6.4.

Modelos com Funções Quadráticas

As funções quadráticas também são usadas com bastante freqüência em economia aplicada para capturar efeitos marginais crescentes ou decrescentes. Seria interessante rever as propriedades das funções quadráticas no Apêndice A, no site da Thomson.

No caso mais simples, y depende de um único fator observado x, mas de uma forma quadrática:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u.$$

Por exemplo, considere $y = salário$ h $x = exper$. Como discutimos no Capítulo 3, esse modelo não se enquadra na análise de regressão simples, mas é facilmente trabalhado em regressão múltipla.

É importante lembrar que β_1 não mede a mudança em y em relação a x; não faz sentido manter x^2 fixo quando se altera x. Se escrevermos a equação estimada como

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2, \qquad \textbf{(6.10)}$$

teremos a aproximação

$$\Delta\hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x, \text{ e assim } \Delta\hat{y}/\Delta x \approx \hat{\beta}_1 + 2\hat{\beta}_2 x. \qquad \textbf{(6.11)}$$

Isso nos mostra que a inclinação da relação entre x e y depende do valor de x; a inclinação estimada é $\hat{\beta}_1 + 2\hat{\beta}_2 x$. Se inserirmos $x = 0$, veremos que $\hat{\beta}_1$ pode ser interpretado como a inclinação aproximada na alteração de $x = 0$ para $x = 1$. Para outras mudanças no valor de x, o segundo termo, $2\hat{\beta}_2 x$, deve ser levado em conta.

Se estivermos interessados em somente computar a mudança prevista em y dado um valor inicial de x e uma mudança de x, poderíamos usar (6.10) diretamente: não há nenhuma razão para usar cálculos de aproximação. Porém, normalmente estamos mais interessados em resumir rapidamente o efeito de x em y, e a interpretação de $\hat{\beta}_1$ e de $\hat{\beta}_2$ na equação (6.11) fornece esse resumo. Em geral, podemos

inserir o valor médio de x na amostra, ou outros valores de interesse, como a mediana ou os valores dos quartis inferior ou superior de x.

Em muitas aplicações, $\hat{\beta}_1$ é positivo, e $\hat{\beta}_2$ é negativo. Por exemplo, utilizando os dados de salários contidos no arquivo WAGE1.RAW, obtemos

$$\begin{gathered} \widehat{salárioh} = 3{,}73 + 0{,}298\, exper - 0{,}0061\, exper^2 \\ (0{,}35)\ (0{,}041) \qquad (0{,}0009) \\ n = 526,\ R^2 = 0{,}093. \end{gathered} \tag{6.12}$$

A equação estimada sugere que *exper* tem um efeito de redução sobre *salárioh*. O primeiro ano de experiência vale aproximadamente 30 centavos de dólar por hora (0,298 dólares). O segundo ano de experiência vale menos [cerca de $0{,}298 - 2(0{,}0061)(1) \approx 0{,}286$, ou 28,6 centavos de dólar, de acordo com a aproximação em (6.11) com $x = 1$]. Aumentando de 10 para 11 os anos de experiência, a previsão de aumento do salário-hora é de cerca de $0{,}298 - 2(0{,}0061)(10) \approx 0{,}176$ ou 17,6 centavos de dólar. E assim por diante.

Quando o coeficiente de x é positivo e o coeficiente de x^2 é negativo, a função quadrática tem um formato parabólico. Sempre existe um valor positivo de x, no qual o efeito de x sobre y é zero; antes desse ponto, x tem um efeito positivo sobre y; após esse ponto, x tem um efeito negativo sobre y. Na prática, pode ser importante saber onde fica esse ponto crítico.

Na equação estimada (6.10) com $\hat{\beta}_1 > 0$ e $\hat{\beta}_2 < 0$, esse ponto crítico (ou o máximo da função) é sempre alcançado na relação entre o coeficiente de x e duas vezes o valor absoluto do coeficiente de x^2:

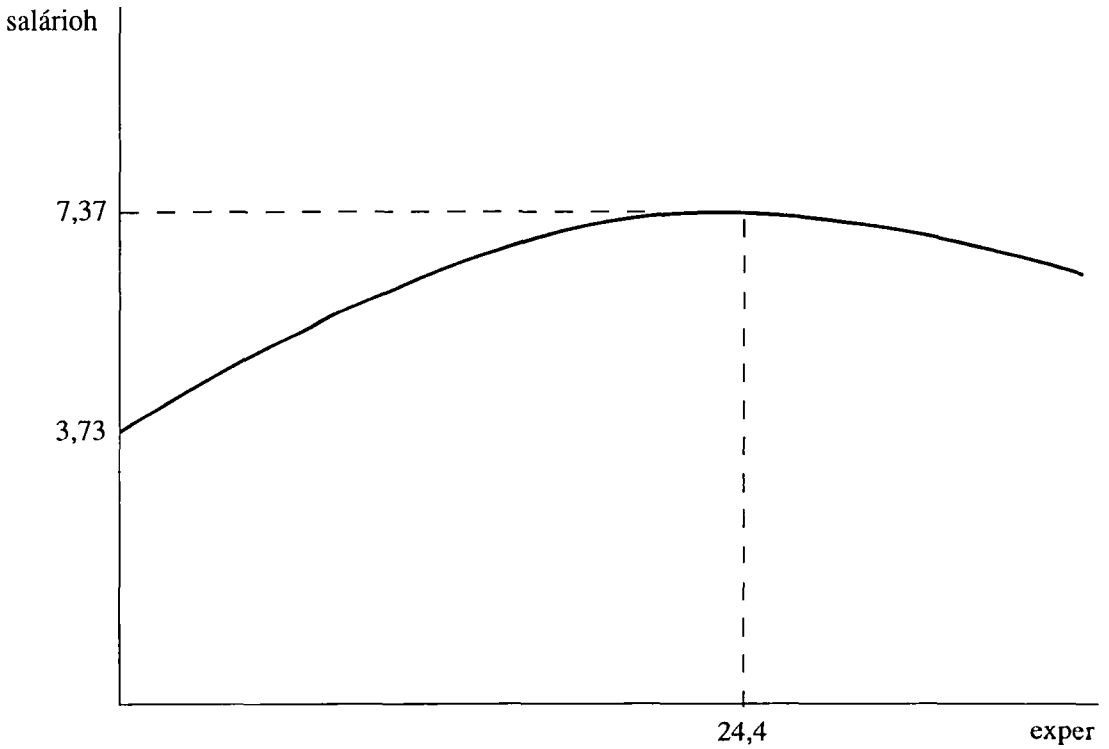
$$x^* = |\hat{\beta}_1/(2\hat{\beta}_2)|. \tag{6.13}$$

No exemplo dos salários, $x^* = exper^*$ é $0{,}298/[2(0{,}0061)] \approx 24{,}4$. (Observe como simplesmente eliminamos o sinal de menos em $-0{,}0061$ ao fazermos esse cálculo.) Esta relação quadrática está ilustrada na Figura 6.1.

Na equação dos salários (6.12), o retorno da experiência passa a ser zero por volta dos 24,4 anos. O que devemos concluir disso? Existem, pelo menos, três explicações possíveis. Primeiro, pode ser que poucas pessoas na amostra tenham mais de 24 anos de experiência, e assim a parte da curva à direita de 24 pode ser ignorada. A conseqüência de se usar uma função quadrática para capturar efeitos decrescentes é que a partir de certo ponto ela acabará fazendo um movimento inverso. Se esse ponto estiver além de uma pequena percentagem das pessoas na amostra, isso não será motivo para grande preocupação. Mas no conjunto de dados do arquivo WAGE1.RAW, cerca de 28% das pessoas na amostra têm mais de 24 anos de experiência; essa é uma percentagem alta demais para se ignorar.

É possível que o retorno de *exper* realmente se torne negativo em algum ponto, mas é difícil acreditar que isso aconteça aos 24 anos de experiência. Uma possibilidade mais provável é que o efeito estimado de *exper* sobre *salárioh* seja viesado, por não termos controlado outros fatores ou porque a relação funcional entre *salárioh* e *exper* na equação (6.12) não está totalmente correta. O Problema 6.9 pede que você explore essa possibilidade controlando a educação, além de usar log(*salárioh*) como a variável dependente.

Figura 6.1
Relação quadrática entre *salárioh* e *exper*.



Quando um modelo tem uma variável dependente na forma logarítmica e uma variável explicativa como uma função quadrática, é necessário certo cuidado para fazer uma boa interpretação. O exemplo seguinte também mostra que a função quadrática pode ter um formato em U, em vez de uma forma parabólica. A forma em U surge na equação (6.10) quando $\hat{\beta}_1$ é negativo e $\hat{\beta}_2$ é positivo; isso captura um efeito crescente de x sobre y.

EXEMPLO 6.2

(Efeitos da Poluição sobre os Preços dos Imóveis)

Modificamos o modelo dos preços dos imóveis do Exemplo 4.5 para incluir um termo quadrático em *comods*:

$$\log(preço) = \beta_0 + \beta_1 \log(oxn) + \beta_2 \log(dist) + \beta_3 comods + \beta_4 comods^2 + \beta_5 razestud + u. \quad (6.14)$$

O modelo estimado utilizando os dados contidos no arquivo HPRICE2.RAW é

$$\log(\hat{preço}) = 13,39 - 0,902 \log(oxn) - 0,087 \log(dist)$$
$$(0,57) \quad (0,115) \qquad (0,043)$$

EXEMPLO 6.2 (continuação)

$$- 0{,}545\ comods + 0{,}062\ comods^2 - 0{,}48\ razestud$$

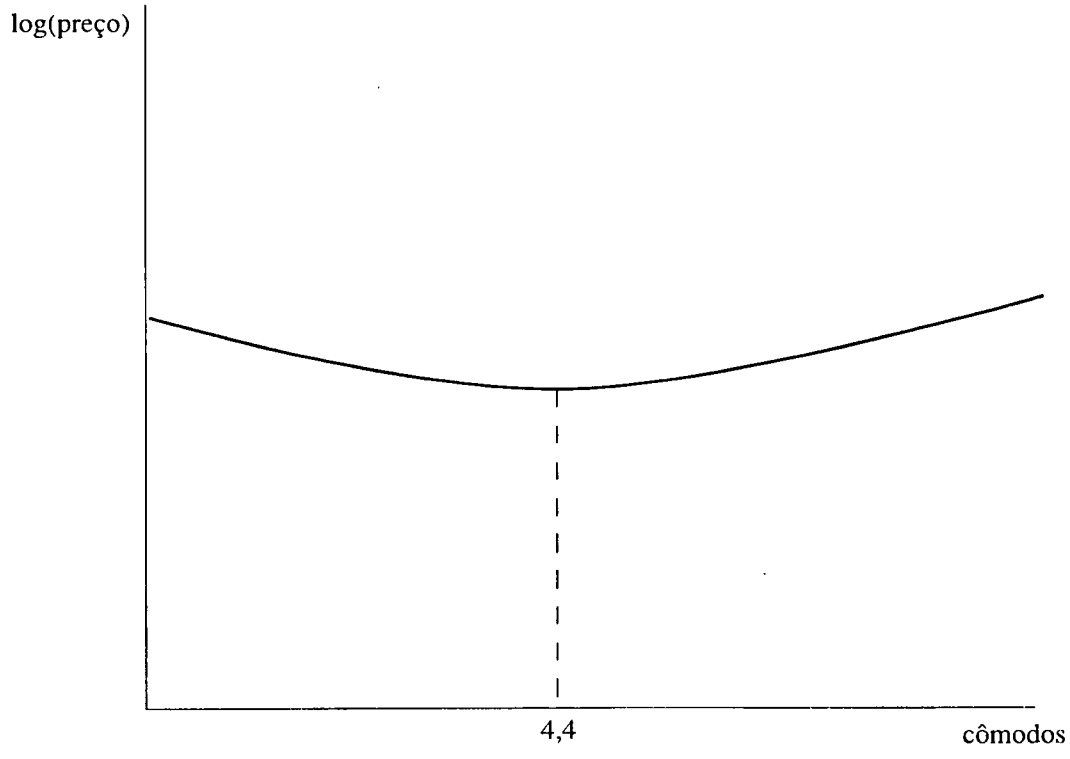
$$(0{,}165) \qquad (0{,}013) \qquad (0{,}006)$$

$$n = 506,\ R^2 = 0{,}603.$$

O termo quadrático $comods^2$ tem uma estatística t em torno de 4,77 e, portanto, é estatisticamente bastante significante. Mas o que é possível afirmar sobre a interpretação do efeito de *comods* sobre log(*preço*)? Inicialmente, o efeito parece ser estranho. Como o coeficiente de *comods* é negativo e o coeficiente de $comods^2$ é positivo, a equação literalmente sugere que, com valores baixos de *comods*, um cômodo adicional tem um efeito *negativo* sobre log(*preço*). Em algum ponto, o efeito se torna positivo, e a forma quadrática significa que a semi-elasticidade de *preço* em relação a *comods* cresce na mesma proporção do crescimento de *comods*. Esta situação é mostrada na Figura 6.2.

Obtemos o valor do ponto crítico de *comods* usando a equação (6.13) (embora $\hat{\beta}_1$ seja negativo e $\hat{\beta}_2$ seja positivo). O valor absoluto do coeficiente de *comods*, 0,545, dividido pelo dobro do coeficiente de $comods^2$, 0,062, resulta em $comods^* = 0{,}545/[2(0{,}062)] \approx 4{,}4$; este ponto está marcado na Figura 6.2.

Figura 6.2
log(*preço*) como uma função quadrática de *comods*.



EXEMPLO 6.2 (continuação)

Será que podemos acreditar que se iniciarmos com três cômodos e aumentarmos para quatro isso efetivamente reduzirá o valor esperado do imóvel? Provavelmente não. Acontece que somente cinco das 506 comunidades na amostra possuem imóveis com média de 4,4 cômodos ou menos, cerca de 1% da amostragem. Isso é tão pequeno que a função quadrática à esquerda de 4,4 pode, para fins práticos, ser ignorada. À direita de 4,4, vemos que a adição de outro cômodo tem um efeito crescente na mudança percentual no preço:

$$\Delta\log(\hat{preço}) \approx \{[-0,545 + 2(0,062)]comods\}\Delta comods$$

e assim

$$\%\Delta\hat{preço} \approx 100\{[-0,545 + 2(0,062)]comods\}\Delta comods$$
$$= (-54,5 + 12,4comods)\,\Delta comods.$$

Portanto, um aumento em *comods*, digamos de cinco para seis, aumenta o preço em aproximadamente $-54,5 + 12,4(5) = 7,5\%$; o aumento de seis para sete aumenta o preço em aproximadamente $-54,5 + 12,4(6) = 19,9\%$. Esse é um efeito crescente bastante forte.

Existem muitas outras possibilidades de usar funções quadráticas juntamente com logaritmos. Por exemplo, uma extensão de (6.14) que permita uma elasticidade não-constante entre *preço* e *oxn* é

$$\log(preço) = \beta_0 + \beta_1\log(oxn) + \beta_2[\log(oxn)]^2 + \beta_3 crime + \beta_4 comods + \beta_5 comods^2 + \beta_6 razestud + u. \quad \textbf{(6.15)}$$

Se $\beta_2 = 0$, β_1 será a elasticidade do *preço* em relação a *oxn*. Caso contrário, essa elasticidade dependerá do nível de *oxn*. Para verificar isso, podemos combinar os argumentos dos efeitos parciais nos modelos quadrático e logarítmico para mostrar que

$$\%\Delta preço \approx [\beta_1 + 2\beta_2\log(oxn)]\%\Delta oxn; \quad \textbf{(6.16)}$$

portanto, a elasticidade do *preço* em relação a *oxn* é $\beta_1 + 2\beta_2\log(oxn)$, de forma que ela depende de $\log(oxn)$.

Finalmente, outros termos polinomiais podem ser incluídos nos modelos de regressão. Certamente a função quadrática é vista com mais freqüência, mas um termo cúbico ou até de quarta potência aparece de vez em quando. Uma forma funcional freqüentemente razoável de uma função de custo total é

$$custo = \beta_0 + \beta_1 quantidade + \beta_2 quantidade^2 + \beta_3 quantidade^3 + u.$$

Não é complicado estimar este modelo. A interpretação dos parâmetros é mais trabalhosa (embora objetiva com o uso de cálculo infinitesimal); não estudaremos este modelo com mais detalhes.

Modelos com Termos de Interação

Algumas vezes, é natural que o efeito parcial, a elasticidade, ou a semi-elasticidade da variável dependente, em relação a uma variável explicativa, dependa da magnitude de *outra* variável explicativa. Por exemplo, no modelo

$$preço = \beta_0 + \beta_1 arquad + \beta_2 qtdorm + \beta_3 arquad \cdot qtdorm + \beta_4 banhos + u,$$

o efeito parcial de *qtdorm* sobre *preço* (mantendo fixas todas as outras variáveis) é

$$\frac{\Delta preço}{\Delta qtdorm} = \beta_2 + \beta_3 arquad. \qquad \textbf{(6.17)}$$

Se $\beta_3 > 0$, (6.17) sugere que um quarto a mais produz um aumento maior no preço dos imóveis maiores. Em outras palavras, existe um **efeito de interação** entre a área do imóvel e o número de quartos. Ao resumirmos o efeito de *qtdorm* sobre *preço*, devemos avaliar (6.17) quanto aos valores de interesse de *arquad*, como o valor médio, ou os quartis inferior ou superior na amostra. Se β_3 é zero ou não, é algo que podemos verificar facilmente.

Pode ser complicado interpretar os parâmetros das variáveis originais quando incluímos um termo de interação. Por exemplo, na equação anterior sobre preços de imóveis, a equação (6.17) mostra que β_2 é o efeito de *qtdorms* sobre *preço* para um preço com zero de área construída! Esse efeito, obviamente, não é muito interessante. Em vez disso, devemos ser cuidadosos ao inserirmos valores de interesse da área do imóvel, como o valor médio ou a mediana da amostra, na versão estimada da equação (6.17).

Freqüentemente, é vantajoso reparametrizar um modelo para que os coeficientes das variáveis originais tenham significados interessantes. Considere um modelo com duas variáveis explicativas e uma interação:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u.$$

Como acabamos de mencionar, β_2 é o efeito parcial de x_2 quando $x_1 = 0$. Muitas vezes, isso não é de interesse. Em vez disso, podemos reparametrizar o modelo como

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u,$$

onde μ_1 é a média populacional de x_1 e μ_2 é a média populacional de x_2. Podemos facilmente ver que agora o coeficiente de x_2, δ_2, é o efeito parcial de x_2 sobre y no valor médio de x_1. (Multiplicando a interação na segunda equação e comparando os coeficientes, podemos mostrar com facilidade que $\delta_2 = \beta_2 + \beta_3 \mu_1$. O parâmetro δ_1 tem uma interpretação semelhante.) Portanto, se subtrairmos as médias das variáveis — na prática, elas seriam, tipicamente, as médias da amostra — antes de criarmos o termo de interação, os coeficientes das variáveis originais terão uma interpretação útil. E mais, obteremos imediatamente os erros-padrão dos efeitos parciais ao nível dos valores médios. Nada nos impede de substituir μ_1 ou μ_2 por outros valores das variáveis explicativas que possam ser de interesse. O exemplo seguinte ilustra como podemos usar os termos de interação.

EXEMPLO 6.3

(Efeitos da Freqüência Escolar no Desempenho de Exames Finais)

Um modelo para explicar o resultado padronizado de um exame final (*respad*) em termos da taxa de freqüência escolar, da nota média anterior* ao curso superior até o penúltimo semestre, e da nota do teste de avaliação de conhecimentos para ingresso em curso superior é

$$respad = \beta_0 + \beta_1 taxafreq + \beta_2 nmgradp + \beta_3 tac + \beta_4 nmgradp^2 + \beta_5 tac^2 + \beta_6 nmgradp \cdot taxafreq + u. \quad \textbf{(6.18)}$$

(Utilizamos o resultado padronizado do exame pelos motivos discutidos na Seção 6.1: é mais fácil interpretar o desempenho de um aluno em relação ao restante da classe.) Além dos termos quadráticos em *nmgradp* e *tac*, este modelo inclui uma interação entre *nmgradp* e *taxafreq*. A idéia é que a freqüência às aulas pode ter um efeito diferente nos alunos que tiveram desempenhos diferentes no passado, como medido por *nmgradp*. Estamos interessados nos efeitos da freqüência sobre as notas do exame final: $\Delta respad/\Delta taxafreq = \beta_1 + \beta_6 nmgradp$.

Usando as 680 observações do arquivo ATTEND.RAW para estudantes da área de economia, a equação estimada é

$$\widehat{respad} = \underset{(1,36)}{2,05} - \underset{(0,0102)}{0,0067}\, taxafreq - \underset{(0,48)}{1,63}\, nmgradp - \underset{(0,098)}{0,128}\, tac + \underset{(0,101)}{0,296}\, nmgradp^2 + \underset{(0,0022)}{0,0045}\, tac^2 + \underset{(0,0043)}{0,0056}\, nmgradp \cdot taxafreq \quad \textbf{(6.19)}$$

$$n = 680,\ R^2 = 0,229,\ \overline{R}^2 = 0,222.$$

Devemos interpretar essa equação com extremo cuidado. Se simplesmente olharmos o coeficiente de *taxafreq*, concluiremos de forma errônea que a freqüência tem um efeito negativo na nota do exame final. Porém, esse coeficiente supostamente mede o efeito quando *nmgradp* = 0, o que não é interessante (nessa amostra, a menor nota média do ensino médio é cerca de 0,86). Também devemos ter cuidado para não examinarmos separadamente as estimativas de β_1 e β_6 e concluirmos que, como cada estatística *t* é não significante, não podemos rejeitar H_0: $\beta_1 = 0, \beta_6 = 0$. Aliás, o *p*-valor do teste *F* dessa hipótese conjunta é 0,014, de modo que com certeza rejeitamos H_0 ao nível de 5%. Este é um bom exemplo de como o exame em separado de estatísticas *t*, quando estamos testando uma hipótese conjunta, pode nos levar a equívocos.

Como devemos estimar o efeito parcial de *taxafreq* sobre *respad*? Devemos inserir valores de interesse de *nmgradp* para obter o efeito parcial. O valor médio de *nmgradp* na amostra é 2,59, de modo que nesse valor médio o efeito de *taxafreq* sobre *respad* é $-0,0067 + 0,0056(2,59) \approx 0,0078$. Qual o significado disso? Como *taxafreq* é medida como um percentual, isso significa que um aumento de dez pontos percentuais em *taxafreq* aumenta *respad* em 0,078 desvios-padrão da nota média do exame final.

Como podemos dizer se a estimativa 0,0078 é estatisticamente diferente de zero? Temos que computar novamente a regressão, substituindo *nmgradp* · *taxafreq* por (*nmgradp* − 2,59) · *taxafreq*. Isso produz, como o novo coeficiente de *taxafreq*, o efeito estimado quando *nmgradp* = 2,59, juntamente com seu erro-padrão; nada mais é alterado na regressão. (Descrevemos esse mecanismo na Seção 4.4.) A execução dessa nova regressão fornece o erro-padrão de $\hat{\beta}_1 + \hat{\beta}_6(2,59) = 0,0078$ como 0,0026, o que produz

EXEMPLO 6.3 (continuação)

$t = 0{,}0078/0{,}026 = 3$. Portanto, na *nmgradp* média, concluímos que a taxa de freqüência às aulas tem um efeito positivo estatisticamente significante nas notas do exame final.

É ainda mais complicado encontrar o efeito de *nmgradp* sobre *respad*, devido ao termo quadrático $nmgradp^2$. Para encontrar o efeito no valor médio de *nmgradp* e na taxa média de freqüência, 0,82, substituímos $nmgradp^2$ por $(nmgradp - 2{,}59)^2$ e $nmgradp \cdot taxafreq$ por $nmgradp \cdot (taxafreq - 0{,}82)$. O coeficiente de *nmgradp* se tornará o efeito parcial nos valores médios e obteremos seu erro-padrão. (Veja o Problema 6.14.)

Se adicionarmos o termo $\beta_7 tac \cdot taxafreq$ à equação (6.18), qual será o efeito parcial de *taxafreq* sobre *respad*?

6.3 UM POUCO MAIS SOBRE O GRAU DE AJUSTE E A SELEÇÃO DE REGRESSORES

Até agora, não dedicamos muita atenção ao tamanho do R^2 na avaliação de nossos modelos de regressão, porque estudantes iniciantes tendem a colocar muito peso no R-quadrado. Como em breve veremos, a seleção de um conjunto de variáveis explicativas com base no tamanho do R-quadrado pode levar a modelos absurdos. No Capítulo 10 descobriremos que R-quadrados obtidos de regressões de séries temporais podem ser artificialmente altos e podem resultar em conclusões enganosas.

Nada nas hipóteses do modelo linear clássico exige que o R^2 esteja acima de qualquer valor em particular; o R^2 é simplesmente uma estimativa do quanto da variação em y é explicado por $x_1, x_2, \ldots, x_k$ na população. Vimos várias regressões que tinham R-quadrados bastante pequenos. Embora isso signifique que não tenhamos avaliado vários fatores que afetam y, isso não quer dizer que os fatores em u sejam correlacionados com as variáveis independentes. A hipótese de média condicional zero RLM.3 é a que determina se obteremos estimadores não-viesados dos efeitos *ceteris paribus* das variáveis independentes, e o tamanho do R-quadrado não tem influência direta nisso.

Um R-quadrado pequeno sugere que a variância do erro é grande em relação à variância de y, o que significa que podemos ter muito trabalho para estimar β_j com precisão. Porém, lembre-se: vimos na Seção 3.4 que uma variância grande do erro pode ser compensada por uma amostra de tamanho grande: se tivermos dados suficientes, podemos ter condições de estimar com precisão os efeitos parciais, mesmo que não tenhamos controlado muitos dos fatores não-observados. Se podemos ou não obter estimativas suficientemente precisas, depende da aplicação que estamos pesquisando. Por exemplo, suponha que alguns alunos ingressantes de uma grande universidade recebam, aleatoriamente, subsídios para a compra de computadores. Se o montante do subsídio for determinado de forma realmente aleatória, podemos estimar o efeito *ceteris paribus* do montante do subsídio sobre a nota de aproveitamento dos alunos, com o uso de uma análise de regressão simples. (Devido à atribuição aleatória, todos os outros fatores que afetam a nota de aproveitamento seriam não-correlacionados com o montante do subsídio.) Parece provável que o montante de subsídio explique pouco da variação na nota de aproveitamento, de modo que o R-quadrado de tal regressão provavelmente será muito pequeno. Mas se tivermos uma amostra de grande tamanho, ainda poderemos ter condições de obter uma estimativa razoavelmente precisa do efeito do subsídio.

Lembre-se, porém, de que a *mudança* relativa no R-quadrado, quando variáveis são adicionadas à equação, é muito útil: a estatística F em (4.41) para testar a significância conjunta depende de forma crucial da diferença nos R-quadrados entre o modelo sem restrições e o modelo restrito.

O *R*-Quadrado Ajustado

A maioria dos programas econométricos registra, juntamente com o R-quadrado, uma estatística chamada ***R*-quadrado ajustado**. Como o R-quadrado ajustado é descrito em muitos trabalhos aplicados, e como ele tem algumas características úteis, trataremos dele nesta subseção.

Para verificar como o R-quadrado usual pode ser ajustado, é útil escrevê-lo como

$$R^2 = 1 - (\text{SQR}/n)/(\text{SQT}/n), \tag{6.20}$$

onde SQR é a soma dos resíduos quadrados e SQT é a soma dos quadrados total; comparada com a equação (3.28), tudo o que fizemos foi dividir tanto SQR como SQT por n. Essa expressão revela o que R^2 está realmente estimando. Defina σ_y^2 como a variância populacional de y e faça com que σ_u^2 represente a variância populacional do termo erro, u. (Até agora temos usado σ^2 para representar σ_u^2, mas é vantajoso ser mais específico neste caso.) O ***R*-quadrado da população** é definido como $1 - \sigma_u^2/\sigma_y^2$; essa é a proporção da variação em y na população explicada pelas variáveis independentes. Isso é o que, supostamente, R^2 deve estar estimando.

O R^2 estima σ_u^2 por SQR/n, que sabemos ser viesado. Então, por que não substituir SQR/n por SQR/$(n - k - 1)$? Além disso, podemos usar SQT/$(n - 1)$ em lugar de SQT/n, já que o primeiro é o estimador não-viesado de σ_y^2. Usando esses estimadores, chegamos ao R-quadrado ajustado:

$$\begin{aligned}\bar{R}^2 &= 1 - [\text{SQR}/(n - k - 1)]/[\text{SQT}/(n - 1)] \\ &= 1 - \hat{\sigma}^2/[\text{SQT}/(n - 1)],\end{aligned} \tag{6.21}$$

já que $\hat{\sigma}^2 = \text{SQR}/(n - k - 1)$. Devido à notação usada para representar o R-quadrado ajustado, ele é, algumas vezes, chamado de *R-barra-quadrado*.

O R-quadrado ajustado algumas vezes é chamado de *R-quadrado corrigido*, mas esse não é um bom nome, pois sugere que $\bar{R}^2$ é de alguma forma melhor que R^2 como um estimador do R-quadrado da população. Infelizmente, $\bar{R}^2$ não é reconhecido, de forma geral, como um melhor estimador. É tentador imaginar que $\bar{R}^2$ corrige o viés de R^2 na estimativa do R-quadrado da população, mas ele não faz isso: a razão de dois estimadores não-viesados não é um estimador não-viesado.

O ponto mais atraente do $\bar{R}^2$ é que ele impõe uma penalidade à inclusão de variáveis independentes adicionais em um modelo. Sabemos que R^2 nunca pode diminuir quando uma nova variável independente é incluída em uma equação de regressão: isso ocorre porque SQR nunca aumenta (e normalmente diminui) quando novas variáveis independentes são adicionadas. Mas a fórmula do $\bar{R}^2$ mostra que ele depende explicitamente de k, o número de variáveis independentes. Se uma variável independente for adicionada a uma regressão, SQR diminui, mas o mesmo acontece com os gl na regressão, $n - k - 1$. Portanto, SQR/$(n - k - 1)$ pode aumentar ou diminuir quando uma nova variável independente é adicionada a uma regressão.

Um fato algébrico interessante é o seguinte: se adicionarmos uma nova variável independente a uma equação de regressão, $\bar{R}^2$ aumenta se, e somente se, a estatística t da nova variável for maior que um em valor absoluto. (Uma extensão disto é que $\bar{R}^2$ aumenta quando um grupo de variáveis é adicionado a uma regressão se, e somente se, a estatística F da significância conjunta das novas variáveis for maior

que a unidade.) Assim, vemos imediatamente que usar o $\bar{R}^2$ para decidir se determinada variável independente (ou conjunto de variáveis) pertence a um modelo nos fornece uma resposta diferente daquelas fornecidas pelos testes usuais t ou F (já que uma estatística t ou F igual à unidade não é estatisticamente significante aos níveis tradicionais de significância).

Algumas vezes é útil termos uma fórmula do $\bar{R}^2$ em termos de R^2. A álgebra simples mostra que

$$\bar{R}^2 = 1 - (1 - R^2)(n - 1)/(n - k - 1). \qquad \textbf{(6.22)}$$

Por exemplo, se $R^2 = 0{,}30$, $n = 51$, e $k = 10$, então $\bar{R}^2 = 1 - 0{,}70(50)/40 = 0{,}125$. Assim, para n pequeno e k grande, $\bar{R}^2$ pode estar substancialmente abaixo de R^2. De fato, se o R-quadrado normal for pequeno, e $n - k - 1$ for pequeno, $\bar{R}^2$ pode, na realidade, ser negativo! Por exemplo, podemos considerar $R^2 = 0{,}10$, $n = 51$, e $k = 10$ para verificar que $\bar{R}^2 = -0{,}125$. Um $\bar{R}^2$ negativo indica uma adaptação muito pobre do modelo relativamente ao número de graus de liberdade.

O R-quadrado ajustado algumas vezes é descrito junto com o R-quadrado habitual em regressões, e algumas vezes o $\bar{R}^2$ é descrito em lugar do R^2. É importante lembrar que é o R^2, e não o $\bar{R}^2$, que aparece na estatística F em (4.41). A mesma fórmula com $\bar{R}^2_r$ e $\bar{R}^2_{ir}$ *não* é válida.

O Uso do *R*-quadrado Ajustado para a Escolha entre Modelos Não-Aninhados

Na Seção 4.5 aprendemos como calcular uma estatística F para testar a significância conjunta de um grupo de variáveis: isso nos possibilita decidir, em um nível particular de significância, se pelo menos uma variável no grupo afeta a variável dependente. Esse teste não nos permite decidir *qual* das variáveis tem um efeito. Em alguns casos, queremos escolher um modelo sem variáveis independentes redundantes, e o R-quadrado ajustado pode nos ajudar nessa tarefa.

No exemplo dos salários dos jogadores da principal liga de beisebol na Seção 4.5, vimos que nem *hrunano* nem *rebrunano* eram individualmente significantes. Essas duas variáveis são altamente correlacionadas, de modo que podemos querer optar entre os modelos

$$\log(salário) = \beta_0 + \beta_1 anos + \beta_2 jogosano + \beta_3 rebmed + \beta_4 hrunano + u$$

e

$$\log(salário) = \beta_0 + \beta_1 anos + \beta_2 jogosano + \beta_3 rebmed + \beta_4 rebrunano + u.$$

Esses dois exemplos são **modelos não-aninhados**, pois nenhuma equação é um caso especial da outra. A estatística F, que estudamos no Capítulo 4, nos permite testar somente modelos *aninhados*: um modelo (o modelo restrito) é um caso especial do outro modelo (o modelo sem restrições). Veja as equações (4.32) e (4.28) para exemplos dos modelos restritos e sem restrições. Uma possibilidade é criar um modelo combinado que contenha *todas* as variáveis explicativas dos modelos originais e depois testar cada modelo contra o modelo geral usando o teste F. O problema deste processo é que ambos os modelos poderão ser rejeitados, ou nenhum modelo poderá ser rejeitado (como acontece com o exemplo dos salários dos jogadores da principal liga de beisebol na Seção 4.5). Assim, esse processo nem sempre fornece uma maneira de fazermos a distinção entre modelos com regressores não-aninhados.

Na regressão dos salários dos jogadores de beisebol, o $\overline{R}^2$ da regressão contendo *hrunano* é 0,6211 e o $\overline{R}^2$ da regressão contendo *rebrunano* é 0,6226. Portanto, com base no R-quadrado ajustado, existe uma preferência muito pequena para o modelo com *rebrunano*. Mas a diferença, na prática, é muito pequena, e podemos obter uma resposta diferente controlando algumas das variáveis no Problema 4.16. (Como ambos os modelos não-aninhados contêm cinco parâmetros, o R-quadrado habitual pode ser usado para fornecer a mesma conclusão.)

A comparação dos $\overline{R}^2$ para optarmos entre os diferentes conjuntos não-aninhados de variáveis independentes pode ser de grande valia quando essas variáveis representam formas funcionais diferentes. Considere dois modelos relacionando a intensidade de P&D às vendas de uma empresa:

$$pdintens = \beta_0 + \beta_1 \log(vendas) + u. \tag{6.23}$$

$$pdintens = \beta_0 + \beta_1 vendas + \beta_2 vendas^2 + u. \tag{6.24}$$

O primeiro modelo captura um rendimento decrescente pela inclusão de *vendas* na forma logarítmica; o segundo modelo faz isso com o uso de um termo quadrático. Assim, o segundo modelo contém um parâmetro a mais que o primeiro.

Quando as equações são estimadas usando as 32 observações das empresas de produtos químicos do arquivo RDCHEM.RAW, o R^2 é 0,061, e o R^2 da equação (6.24) é 0,148. Portanto, parece que a função quadrática faz um ajuste muito melhor. Mas uma comparação dos R-quadrados habituais com o primeiro modelo é injusta, porque ele contém um parâmetro a menos que a equação (6.24). Isto é, (6.23) é um modelo mais parcimonioso que (6.24).

Tudo o mais igual, modelos mais simples são melhores. Como o R-quadrado habitual não penaliza modelos mais complicados, é melhor usar o $\overline{R}^2$. O $\overline{R}^2$ de (6.23) é 0,030, enquanto o $\overline{R}^2$ de (6.24) é 0,090. Portanto, mesmo depois dos ajustes das diferenças nos graus de liberdade, o modelo quadrático é o melhor. O modelo quadrático também é o preferido quando margens de lucro são incluídas em cada regressão.

Existe uma limitação importante no uso do $\overline{R}^2$ para escolher entre modelos não-aninhados: não podemos usá-lo para a escolha entre diferentes formas funcionais da variável dependente. Isso é uma pena, pois muitas vezes queremos saber se y ou $\log(y)$ (ou talvez alguma outra transformação) deve ser usada como a variável dependente com base no grau de ajuste. Mas nem o R^2 nem o $\overline{R}^2$ podem ser usados para esse fim. A razão é simples: esses R-quadrados medem a proporção explicada do total da variação de qualquer variável dependente que estejamos usando na regressão, e diferentes funções da variável dependente terão diferentes montantes de variação a ser explicadas. Por exemplo, as variações totais em y e $\log(y)$ não são as mesmas. A comparação dos R-quadrados ajustados dessas regressões com essas diferentes formas das variáveis dependentes não nos dá nenhuma informação sobre qual modelo se adapta melhor; eles estimam duas variáveis dependentes separadas.

Explique por que escolher um modelo maximizando $\overline{R}^2$ ou minimizando $\hat{\sigma}$ (o erro-padrão da regressão) é a mesma coisa.

EXEMPLO 6.4

(Remuneração de Diretores Executivos e Desempenho de Empresas)

Considere dois modelos estimados relacionando a remuneração de diretores executivos ao desempenho de empresas:

$$\widehat{salário} = 830{,}63 + 0{,}0163\ vendas + 19{,}63\ rma \quad \textbf{(6.25)}$$
$$(223{,}90)\ (0{,}0089)\ (11{,}08)$$
$$n = 209, R^2 = 0{,}029, \overline{R}^2 = 0{,}20$$

e

$$\widehat{lsalário} = 4{,}36 + 0{,}275\ lvendas + 0{,}0179\ rma \quad \textbf{(6.26)}$$
$$(0{,}29)\ (0{,}033)\ (0{,}0040)$$
$$n = 209, R^2 = 0{,}282, \overline{R}^2 = 0{,}275,$$

onde *rma* é o retorno das ações, discutido no Capítulo 2. Para simplificar, *lsalário* e *lvendas* representam os logs naturais de *salário* e *vendas*. Já sabemos como interpretar essas diferentes equações estimadas. Mas podemos dizer se um modelo ajusta os dados melhor que o outro?

O R-quadrado da equação (6.25) mostra que *vendas* e *rma* explicam somente cerca de 2,9% da variação do salário dos diretores executivos na amostra. Tanto *vendas* como *rma* têm significância estatística marginal.

A equação (6.26) mostra que log(*vendas*) e *rma* explicam cerca de 28,2% da variação do log(salário). Em termos de grau de ajuste, esse R-quadrado bem mais alto parece sugerir que o modelo (6.26) é bem melhor, mas esse não é necessariamente o caso. A soma dos quadrados total de *salário* na amostra é 391.732.982, enquanto a soma dos quadrados total de log(*salário*) é somente 66,72. Assim, há muito menos variação em log(*salário*) que precisa ser explicada.

Neste ponto, podemos usar outros recursos além do R^2 e do $\overline{R}^2$ para optar entre esses modelos. Por exemplo, log(*vendas*) e *rma* são muito mais significantes, estatisticamente, em (6.26) do que são *vendas* e *rma* em (6.25), e os coeficientes em (6.26) provavelmente são de maior interesse. Para termos certeza, porém, precisaremos fazer uma comparação válida dos graus de ajuste.

Na Seção 6.4 forneceremos um indicador que efetivamente nos permita comparar modelos nos quais y aparece tanto na forma em nível como na forma logarítmica.

O Controle de muitos Fatores na Análise de Regressão

Em muitos dos exemplos que tratamos, e certamente em nossa discussão sobre o viés de variáveis omitidas no Capítulo 3, temos nos preocupado com a omissão de fatores importantes em modelos que possam estar correlacionados com as variáveis independentes. Também é possível controlarmos grande quantidade de variáveis em uma análise de regressão.

Se enfatizarmos exageradamente o grau de ajuste, estaremos nos propondo a controlar fatores em um modelo de regressão que não deveriam ser controlados. Para evitar este equívoco, precisamos nos lembrar da interpretação *ceteris paribus* de modelos de regressão múltipla.

Para ilustrar esse problema, suponha que estejamos fazendo um estudo para avaliar o impacto dos impostos estaduais sobre a cerveja em acidentes fatais de trânsito. A idéia é que um imposto mais elevado sobre a cerveja reduzirá o consumo de bebidas alcoólicas e, da mesma forma, o hábito de dirigir sob embriaguez, resultando em menos acidentes fatais de trânsito. Para medirmos o efeito *ceteris paribus* dos impostos sobre esses acidentes, podemos modelar *fatalidades* como uma função de diversos fatores, inclusive o *imposto* sobre a cerveja:

$$fatalidades = \beta_0 + \beta_1 imposto + \beta_2 milhas + \beta_3 percmasc + \beta_4 perc16\_21 + ...,$$

onde *milhas* é o total de milhas dirigidas, *percmasc* é a percentagem masculina da população do Estado, e *perc16_21* é a percentagem da população entre 16 e 21 anos de idade, e assim por diante. Observe que não incluímos uma variável medindo o consumo *per capita* de cerveja. Estaremos cometendo um erro de variáveis omitidas? A resposta é não. Se controlarmos o consumo de cerveja nessa equação, de que forma o imposto sobre cerveja afetará as fatalidades no trânsito? Na equação

$$fatalidades = \beta_0 + \beta_1 imposto + \beta_2 conscerv + ...,$$

β_1 mede a diferença nas fatalidades devido ao aumento de um ponto percentual no imposto, mantendo *conscerv* fixo. É difícil de entender por que isso seria de interesse. Não deveríamos controlar as diferenças de *conscerv* entre os Estados, a menos que quiséssemos verificar algum tipo de efeito indireto do imposto sobre a cerveja. Outros fatores, como a distribuição por sexo e idade, deveriam ser controlados.

A questão de decidir se devemos ou não controlar certos fatores nem sempre é bem definida. Por exemplo, Betts (1995) estuda o efeito da qualidade do ensino médio sobre a renda subseqüente. Ele salienta que, se qualidade melhor de ensino resulta em mais educação, então controlar a educação na regressão juntamente com avaliação da qualidade subestimará o retorno da qualidade. Betts faz a análise com e sem anos de escolaridade na equação para obter uma gama de efeitos estimados da qualidade de ensino.

Para verificar claramente como a ênfase em R-quadrados altos pode criar problemas, considere o exemplo do preço dos imóveis da Seção 4.5 que ilustra a verificação de múltiplas hipóteses. Naquele caso, queríamos verificar a racionalidade da avaliação dos preços dos imóveis. Fizemos a regressão de log(*preço*) sobre log(*aval*), log(*tamterr*), log(*arquad*), e *qtdorm* e verificamos se as três últimas variáveis tinham coeficientes populacionais zero enquanto log(*aval*) tinha um coeficiente unitário. Entretanto, se quisermos estimar um modelo de preço hedônico, como no Exemplo 4.8, onde os valores marginais de várias características dos imóveis são obtidos? Devemos incluir log(*aval*) na equação? O R-quadrado ajustado da regressão com log(*aval*) é 0,762, enquanto o R-quadrado ajustado sem ele é 0,630. Com base somente no grau de ajuste, devemos incluir log(*aval*). Contudo, isso será incorreto se nossa meta for determinar os efeitos do tamanho da propriedade, área construída e número de quartos nos valores dos imóveis. A inclusão de log(*aval*) na equação equivale a manter um indicador de valor fixo e indagar quanto a adição de um quarto alterará outro indicador de valor. Essa medida não faz sentido na avaliação das características dos imóveis.

Se lembrarmos que modelos diferentes servem a propósitos diferentes, e nos concentrarmos na interpretação *ceteris paribus* da regressão, não incluiremos os fatores errados em um modelo de regressão.

A Adição de Regressores para Reduzir a Variância do Erro

Acabamos de ver alguns exemplos nos quais certas variáveis independentes não devem ser incluídas em um modelo de regressão, mesmo que elas sejam correlacionadas com a variável dependente. Do Capítulo 3, sabemos que a adição de uma nova variável independente em uma regressão pode exa-

cerbar o problema da multicolinearidade. De outro lado, como estamos retirando algo do termo erro, a adição de uma variável geralmente reduz a variância do erro. De forma geral, não podemos saber que efeito será dominante.

Porém, há um caso que é óbvio: devemos sempre incluir variáveis independentes que afetem y e que sejam *não-correlacionadas* com todas as variáveis independentes de interesse. A razão para esta inclusão é simples: a adição dessa variável não induz multicolinearidade na população (e, portanto, a multicolinearidade na amostra deve ser desprezível), mas reduzirá a variância do erro. Em amostras de tamanho grande, os erros-padrão de todos os estimadores MQO serão reduzidos.

Como exemplo, considere estimar a demanda individual por cerveja como uma função do preço médio da cerveja no município. Pode ser razoável assumir que as características individuais sejam não-correlacionadas com os preços em nível de municípios, e assim uma regressão simples do consumo de cerveja sobre o preço nos municípios seria suficiente para estimar o efeito do preço sobre a demanda individual. Entretanto, é possível obter uma estimativa mais precisa da elasticidade-preço da demanda por cerveja com a inclusão de características individuais, como a idade e o grau de escolaridade. Se esses fatores afetarem a demanda e forem não-correlacionados com o preço, o erro-padrão do coeficiente do preço será menor, pelo menos em amostras grandes.

Como segundo exemplo, considere o subsídio para equipamentos de computação dado no início da Seção 6.3. Se, além da variável do subsídio, controlarmos outros fatores que possam explicar a nota média em curso superior, poderemos provavelmente conseguir uma estimativa mais precisa do efeito do subsídio. Variáveis indicadoras da nota média no ensino médio, a classificação da instituição, as pontuações sat e tac e os antecedentes familiares são bons candidatos. Como os montantes do subsídio são determinados aleatoriamente, todas as variáveis de controle adicionais serão não-correlacionadas com o montante de subsídio; nessa amostra, a multicolinearidade entre o montante do subsídio e as outras variáveis independentes deve ser mínima. Porém, a adição de controles extras pode reduzir significantemente a variância do erro, conduzindo a uma estimativa mais precisa do efeito do subsídio. Lembre-se, neste caso, de que o problema não é a inexistência de viés: obteremos um estimador não-viesado e consistente, quer incluamos ou não as variáveis de desempenho no ensino médio e de antecedentes familiares. O problema está na obtenção de um estimador com uma menor variância amostral.

Infelizmente, casos em que temos informações sobre as variáveis explicativas adicionais que sejam não-correlacionadas com as variáveis explicativas de interesse são raros no campo das ciências sociais. Porém, vale a pena lembrar que, quando essas variáveis estão disponíveis, elas poderão ser incluídas em um modelo para reduzir a variância do erro sem induzir multicolinearidade.

6.4 PREVISÃO E ANÁLISE DE RESÍDUOS

No Capítulo 3 definimos os valores previstos ou estimados do MQO e os resíduos do MQO. As **previsões** certamente são úteis, mas estão sujeitas à variação amostral, já que elas são obtidas com o uso dos estimadores MQO. Assim, nesta seção, mostramos como obter intervalos de confiança de previsões da linha de regressão MQO.

Sabemos, dos capítulos 3 e 4, que os resíduos são usados para obter a soma dos resíduos quadrados e o R-quadrado, de modo que eles são importantes para o grau de ajuste e os testes de hipóteses.

Algumas vezes, os economistas estudam os resíduos de uma observação específica para obter informações sobre os indivíduos (ou empresas, imóveis etc.) na amostra.

Intervalos de Confiança de Previsões

Suponha que tenhamos estimado a equação

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k. \tag{6.27}$$

Quando inserimos valores específicos das variáveis independentes nessa equação, obtemos uma previsão de y, que é uma estimativa do *valor esperado* de y, dados os valores específicos das variáveis explicativas. Para enfatizar, sejam c_1, c_2, ..., c_k valores particulares de cada uma das k variáveis independentes; elas poderão ou não corresponder a um ponto efetivo dos dados em nossa amostra. O parâmetro que gostaríamos de estimar é

$$\begin{aligned} \theta_0 &= \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \ldots + \beta_k c_k \\ &= \mathrm{E}(y \mid x_1 = c_1, x_2 = c_2, \ldots, x_k = c_k). \end{aligned} \tag{6.28}$$

O estimador de θ_0 é

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \ldots + \hat{\beta}_k c_k. \tag{6.29}$$

Na prática, isso é fácil de ser computado. Entretanto, se o que quisermos for um indicador da incerteza nesse valor previsto? É natural construir um intervalo de confiança de θ_0 que seja centrado em $\hat{\theta}_0$.

Para obter um intervalo de confiança de θ_0 precisamos de um erro-padrão de $\hat{\theta}_0$. Então, com um grande gl, poderemos construir um intervalo de confiança de 95% utilizando a regra prática $\hat{\theta}_0 \pm 2 \cdot \mathrm{ep}(\hat{\theta}_0)$. (Como sempre, podemos usar os percentis exatos em uma distribuição t.)

Como obtemos o erro-padrão de $\hat{\theta}_0$? Este é o mesmo problema que encontramos na Seção 4.4: precisamos obter um erro-padrão de uma combinação linear dos estimadores MQO. Aqui, o problema é ainda mais complicado, pois todos os estimadores MQO geralmente aparecem em $\hat{\theta}_0$ (a menos que algum c_j seja zero). No entanto, o mesmo truque que usamos na Seção 4.4 funcionará aqui. Escreva $\beta_0 = \theta_0 - \beta_1 c_1 - \ldots - \beta_k c_k$ e agregue isso à equação

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

para obter

$$y = \theta_0 + \beta_1 (x_1 - c_1) + \beta_2 (x_2 - c_2) + \ldots + \beta_k (x_k - c_k) + u. \tag{6.30}$$

Em outras palavras, subtraímos o valor c_j de cada observação de x_j, e depois computamos a regressão de

$$y_i \text{ sobre } (x_{i1} - c_1), \ldots, (x_{ik} - c_k), \; i = 1, 2, \ldots, n. \tag{6.31}$$

O valor previsto em (6.29) e, mais importante, seu erro-padrão, são obtidos do *intercepto* (ou constante) na regressão (6.31).

Como exemplo, obtemos um intervalo de confiança de uma previsão a partir de uma regressão de nota média em curso superior, das quais usamos informações do ensino médio.

EXEMPLO 6.5

(Intervalo de Confiança de *nmgrad* Previsto)

Utilizando os dados contidos no arquivo GPA2.RAW, obtemos a seguinte equação para prever *nmgrad*:

$$\begin{gathered} \widehat{nmgrad} = 1{,}493 + 0{,}00149\, sat - 0{,}01386\, emperc \\ (0{,}075)\ \ (0{,}00007)\qquad (0{,}00056) \\ - 0{,}06088\, tamclas + 0{,}00546\, tamclas^2 \qquad \textbf{(6.32)} \\ (0{,}1650)\qquad\qquad (0{,}00227) \\ n = 4.137,\ R^2 = 0{,}278,\ \overline{R}^2 = 0{,}277,\ \hat{\sigma} = 0{,}560, \end{gathered}$$

onde apresentamos as estimativas com várias casas decimais para reduzir o erro de arredondamento. Qual a previsão de *nmgrad* quando *sat* = 1.200, *emperc* = 30 e *tamclas* = 5 (o que significaria 500)? Isso é fácil de ser obtido, incorporando esses valores na equação (6.32): $\widehat{nmgrad}$= 2,70 (arredondado para duas casas decimais). Infelizmente, não podemos usar diretamente a equação (6.32) para obter um intervalo de confiança da *nmgrad* esperada com os valores dados das variáveis independentes. Uma maneira simples de obter um intervalo de confiança é definir um novo conjunto de variáveis independentes: *sat0* = sat − 1.200, *emperc0* = *emperc* − 30, *tamclas0* = *tamclas* − 5 e *tamclasquad0* = $tamclas^2$ − 25. Quando fazemos a regressão de *nmgrad* sobre essas novas variáveis independentes, obtemos

$$\begin{gathered} \widehat{nmgrad} = 2{,}700 + 0{,}00149\, sat0 - 0{,}01386\, emperc0 \\ (0{,}020)\ \ (0{,}00007)\qquad (0{,}00056) \\ - 0{,}06088\, tamclas0 + 0{,}00546\, tamclasquad0 \\ (0{,}1650)\qquad\qquad (0{,}00227) \\ n = 4.137,\ R^2 = 0{,}278,\ \overline{R}^2 = 0{,}277,\ \hat{\sigma} = 0{,}560. \end{gathered}$$

A única diferença entre esta regressão e aquela em (6.32) é o intercepto, que é a previsão que queremos, juntamente com seu erro-padrão, 0,020. Não é por acidente que os coeficientes de inclinação, seus erros-padrão, *R*-quadrado etc. são os mesmos de antes; esse fato fornece uma maneira de verificarmos se foram feitas as transformações adequadas. Podemos construir com facilidade um intervalo de confiança de 95% da nota média esperada: 2,70 ± 1,96(0,020) ou em torno de 2,66 a 2,74. Este intervalo de confiança é suficientemente estreito devido ao tamanho bastante grande da amostra.

Como a variância do estimador do intercepto é a menor quando cada variável explicativa tem média amostral zero (veja a Questão 2.5 para o caso da regressão simples), segue da regressão em (6.31) que a variância da previsão nos valores médios de x_j (isto é, $c_j = \bar{x}_j$ para todo j) é a menor. Este resultado não é tão surpreendente, já que o ponto de maior confiabilidade em nossa linha de regressão está próximo ao centro dos dados. Na medida em que os valores de c_j se afastam de $\bar{x}_j$, Var($\hat{y}$) se torna cada vez maior.

O método anterior nos possibilita colocar um intervalo de confiança em torno da estimativa MQO de $E(y|x_1, \ldots, x_k)$ para quaisquer valores das variáveis explicativas. Em outras palavras, obtemos um

intervalo de confiança do valor médio de y da subpopulação com determinado conjunto de covariadas. Entretanto, um intervalo de confiança da média pessoal na subpopulação não é a mesma coisa que um intervalo de confiança de uma unidade particular (indivíduo, família, empresa etc.) da população. Na formação de um intervalo de confiança de um resultado desconhecido de y, devemos avaliar outra fonte muito importante de variação: a variância no erro não observado, que registra nosso desconhecimento dos fatores não observados que afetam y.

Seja y^0 o valor para o qual gostaríamos de construir um intervalo de confiança, que algumas vezes chamamos de **intervalo de previsão**. Por exemplo, y^0 poderia representar uma pessoa ou uma empresa que não esteja em nossa amostra original. Façamos $x_1^0 \ldots , x_k^0$ serem os novos valores das variáveis independentes, que assumimos observar, e u^0 ser o erro não observado. Portanto, temos

$$y^0 = \beta_0 + \beta_1 x_1^0 + \beta_2 x_2^0 + \ldots + \beta_k x_k^0 + u^0. \qquad \textbf{(6.33)}$$

Como antes, nossa melhor previsão de y^0 é o valor esperado de y^0, dadas as variáveis explicativas que estimamos da linha de regressão MQO: $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \hat{\beta}_2 x_2^0 + \ldots + \hat{\beta}_k x_k^0$. O **erro de previsão** com o uso de $\hat{y}^0$ para prever y^0 é

$$\hat{e}^0 = y^0 - \hat{y}^0 = (\beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0) + u^0 - \hat{y}^0. \qquad \textbf{(6.34)}$$

Agora, $E(\hat{y}^0) = E(\hat{\beta}_0) + E(\hat{\beta}_1)x_1^0 + E(\hat{\beta}_2)x_2^0 + \ldots + E(\hat{\beta}_k)x_k^0 = \beta_0 + \beta_1 x_1^0 + \ldots + \beta_k x_k^0$, porque os $\hat{\beta}_j$ são não-viesados. (Como antes, essas expectativas são todas condicionais aos valores amostrais das variáveis independentes.) Como u^0 tem média zero, $E(\hat{e}^0) = 0$. Mostramos que o erro de previsão esperado é zero.

Ao encontrar a variância de $\hat{e}^0$, observe que u^0 é não-correlacionado com cada $\hat{\beta}_j$, porque u^0 é não-correlacionado com os erros na amostra usada para a obtenção de $\hat{\beta}_j$. Pelas propriedades básicas da covariância (veja Apêndice B, no site da Thomson), u^0 e $\hat{y}^0$ são não-correlacionados. Portanto, a **variância do erro de previsão** (condicional a todos os valores das variáveis independentes incluídas na amostra) é a soma das variâncias:

$$\text{Var}(\hat{e}^0) = \text{Var}(\hat{y}^0) + \text{Var}(u^0) = \text{Var}(\hat{y}^0) + \sigma^2, \qquad \textbf{(6.35)}$$

onde $\sigma^2 = \text{Var}(u^0)$ é a variância do erro. Existem duas fontes de variância em $\hat{e}^0$. A primeira é o erro de amostragem em $\hat{y}^0$, que surge por termos estimado β_j. Como cada $\hat{\beta}_j$ tem uma variância proporcional a $1/n$, na qual n é o tamanho da amostra, $\text{Var}(\hat{y}^0)$ é proporcional a $1/n$. Isso significa que, para amostras grandes, $\text{Var}(\hat{y}^0)$ pode ser muito pequena. Em contraposição, σ^2 é a variância do erro na população; ela não muda com o tamanho da amostra. Em muitos exemplos, σ^2 será o termo dominante em (6.35).

Sob as hipóteses do modelo linear clássico, $\hat{\beta}_j$ e u^0 são normalmente distribuídos, e assim $\hat{e}^0$ também é normalmente distribuído (condicional a todos os valores amostrais das variáveis explicativas). Anteriormente, descrevemos como obter um estimador não-viesado de $\text{Var}(\hat{y}^0)$, e obtivemos nosso estimador não-viesado de σ^2 no Capítulo 3. Com o uso desses estimadores, podemos definir o erro-padrão de $\hat{e}^0$ como

$$ep(\hat{e}^0) = \{[ep(\hat{y}^0)]^2 + \hat{\sigma}^2\}^{1/2}. \quad \textbf{(6.36)}$$

Utilizando o mesmo raciocínio para as estatísticas t de $\hat{\beta}_j$, $\hat{e}^0 / ep(\hat{e}^0)$ tem uma distribuição t com $n - (k + 1)$ graus de liberdade. Portanto,

$$P[-t_{0,025} \leq \hat{e}^0 / ep(\hat{e}^0) \leq t_{0,025}] = 0,95,$$

onde $t_{0,025}$ é o 97,5º percentil na distribuição t_{n-k-1}. Para $n - k - 1$ grande, lembre-se de que $t_{0,025} \approx 1,96$. Inserindo $\hat{e}^0 = y^0 - \hat{y}^0$ e fazendo a reordenação, obtemos um intervalo de previsão de 95% para y^0:

$$\hat{y}^0 \pm t_{0,025} \cdot ep(\hat{e}^0); \quad \textbf{(6.37)}$$

como sempre, exceto para gl pequeno, uma boa regra prática é $\hat{y}^0 \pm 2ep(\hat{e}^0)$. Isso é mais amplo que o próprio intervalo de confiança de $\hat{y}^0$, devido a $\hat{\sigma}^2$ em (6.36); normalmente ela é muito mais ampla para refletir os fatores em u^0 que não tenhamos controlado.

EXEMPLO 6.6

(Intervalo de Confiança de Notas Médias Futuras)

Suponha que desejamos um IC de 95% de *nmgrad* futuro de um aluno do ensino médio com *sat* = 1.200, *emperc* = 30 e *tamclas* = 5. No Exemplo 6.5 obtivemos um intervalo de confiança de 95% da *média* da nota média entre todos os alunos com as características particulares *sat* = 1.200, *emperc* = 30 e *tamclas* = 5. Agora, queremos um intervalo de confiança de 95% de qualquer aluno que *especificamente* tenha essas características. O intervalo de previsão de 95% deve registrar a variação na característica individual, não-observada, que afeta o desempenho universitário. Temos tudo que é preciso para obter um IC de *nmgrad*. Sabemos que $ep(\hat{y}^0) = 0,020$ e $\hat{\sigma} = 0,560$ e, portanto, de (6,36), $ep(\hat{e}^0) = [(0,020)^2 + (0,560)^2]^{1/2} \approx 0,560$. Observe o quanto $ep(\hat{y}^0)$ é pequeno em relação a $\hat{\sigma}$: virtualmente, toda a variação em $\hat{e}^0$ vem da variação em u^0. O IC de 95% é $2,70 \pm 1,96(0,560)$ ou está entre 1,60 e 3,80. Este é um intervalo de confiança enorme, e mostra que, com base nos fatores que incluímos na regressão, não podemos definir com clareza a futura nota de graduação de determinado indivíduo. (Em certo sentido, isso é bom, por significar que a classificação no curso médio e o desempenho no teste de aptidão acadêmica não predeterminam o desempenho de alguém na faculdade.) Evidentemente, as características não observadas variam amplamente de um indivíduo para o outro com as mesmas notas no teste de aptidão acadêmica e na classificação no curso médio observadas.

Análise de Resíduos

Algumas vezes, é útil examinar as observações individuais para verificar se o valor efetivo da variável dependente está acima ou abaixo do valor previsto; isto é, examinar os resíduos das observações individuais. Este processo é chamado **análise de resíduos**. Os economistas são conhecidos por examinarem os resíduos de uma regressão para auxiliá-los, por exemplo, na compra de um imóvel.

O exemplo seguinte sobre preços de imóveis ilustra a análise de resíduos. Os preços dos imóveis estão relacionados a várias características observadas do imóvel. Podemos relacionar todas as características que julgarmos importantes, como tamanho, número de quartos, número de banheiros, e assim por diante. Podemos usar uma amostra de imóveis para estimar o relacionamento entre o preço e as características, e terminamos obtendo um valor previsto e um valor real de cada imóvel. Então, podemos construir os resíduos, $\hat{u}_i = y_i - \hat{y}_i$. O imóvel com o maior resíduo negativo é, pelo menos com base nos fatores que estamos controlando, o mais barato em relação às suas características *observadas*. É claro que um preço de venda substancialmente inferior ao seu preço previsto poderia indicar alguma característica indesejável do imóvel que deixamos de avaliar, e que portanto está contido no erro não-observado. Além da obtenção da previsão e do resíduo, também faz sentido computar o intervalo de confiança de qual seria o preço de venda do imóvel no futuro, utilizando o método descrito na equação (6.37).

Utilizando os dados contidos no arquivo HPRICE1.RAW, computamos a regressão de *preço* sobre *tamterr*, *arquad* e *qtdorm*. Na amostra de 88 imóveis, o resíduo mais negativo é −120,206, do 81º imóvel. Portanto, o preço pedido por esse imóvel está US$120.206,00 abaixo de seu preço previsto.

Existem muitos outros usos da análise de resíduos. Uma maneira de classificar as faculdades de direito é fazer a regressão da mediana dos salários iniciais sobre uma variedade de características dos alunos (como a mediana das notas de ingresso nos cursos para novos alunos, a mediana das notas médias de graduação para novos alunos etc.) e obter um valor previsto e um resíduo de cada faculdade de direito. A faculdade de direito com o maior resíduo terá o maior valor agregado previsto. (Naturalmente, ainda existirá muita incerteza sobre como o salário inicial de um indivíduo se compararia com a mediana geral de uma faculdade de direito.) Esses resíduos poderão ser usados juntamente com mensalidades cobradas pelas faculdades de direito para determinarmos o valor mais vantajoso; isso exigirá um desconto apropriado dos ganhos futuros.

A análise de resíduos também tem participação em decisões judiciais. Um artigo publicado no jornal *The New York Times* intitulado "Juiz Diz que Pobreza de Alunos, e Não a Segregação, Prejudica Aproveitamento Escolar" (28.06.95) descreve um importante processo legal. A questão era se o fraco desempenho nos exames padronizados do Distrito Escolar de Hartford, em relação ao desempenho nos distritos vizinhos, era devido à baixa qualidade de ensino nas escolas altamente segregadas. O juiz concluiu que "a disparidade nas notas de aproveitamento escolar não indica que Hartford esteja fazendo um trabalho inadequado ou insuficiente na educação de seus alunos ou que suas escolas sejam deficientes, pois as notas de aproveitamento previstas com base em relevantes fatores socioeconômicos estão próximas dos níveis esperados". Esta conclusão foi, quase com certeza, baseada em uma análise de regressão das notas de aproveitamento médias ou de suas medianas sobre as características socioeconômicas de vários distritos escolares de Connecticut. A conclusão do juiz sugere que, considerando o nível de pobreza dos alunos das escolas de Hartford, as notas de aproveitamento efetivas dos alunos eram semelhantes às previstas em uma análise de regressão: o resíduo de Hartford não era suficientemente negativo para se concluir que as escolas em si mesmas eram responsáveis pelas baixas notas de aproveitamento escolar.

Como você poderia usar a análise de resíduos para determinar quais atores cinematográficos são remunerados em níveis acima da bilheteria dos filmes em que atuam?

Previsão de *y* quando a Variável Dependente é log(*y*)

Como a transformação do log natural é usada com tanta freqüência na variável dependente em economia empírica, dedicamos esta subseção ao problema de prognosticar *y* quando a variável dependente é log(*y*). Como um subproduto, obteremos um indicador de grau de ajuste do modelo log que possa ser comparado com o *R*-quadrado do modelo em nível.

Para obter uma previsão é útil definirmos *logy* = log(*y*); isso realça o fato de que é o log de *y* que será previsto no modelo

$$logy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u. \qquad \textbf{(6.38)}$$

Nesta equação os x_j poderão ser transformações de outras variáveis; por exemplo, poderíamos ter x_1 = log(*vendas*), x_2 = log(*valmerc*), x_3 = *permceo* no exemplo do salário dos diretores executivos.

Dados os estimadores MQO, sabemos como prever *logy* para qualquer valor das variáveis independentes:

$$\hat{logy} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k. \qquad \textbf{(6.39)}$$

Agora, como o exponencial desfaz o log, nossa primeira suposição para prever *y* é simplesmente exponenciar o valor previsto de log(*y*): $\hat{y} = \exp(\hat{logy})$. Isso não funciona; aliás, isso sistematicamente *subestimará* o valor esperado de *y*. De fato, se o modelo (6.38) obedecer às hipóteses do modelo linear clássico RLM.1 até RLM.6, pode ser demonstrado que

$$E(y|\mathbf{x}) = \exp(\sigma^2/2) \cdot \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)$$

onde $\mathbf{x}$ representa as variáveis independentes e σ^2 é a variância de *u*. [Se $u \sim$ Normal(0,σ^2), o valor esperado de exp(*u*) será exp(σ^2/2)]. Esta equação mostra que um ajuste simples é necessário para prevermos *y*:

$$\hat{y} = exp(\hat{\sigma}^2/2)exp(\hat{logy}), \qquad \textbf{(6.40)}$$

onde $\hat{\sigma}^2$ é simplesmente o estimador não-viesado de σ^2. Como $\hat{\sigma}$, o erro-padrão da regressão, é sempre conhecido, a obtenção de valores previstos de *y* será fácil. Como $\hat{\sigma}^2 > 0$, $\exp(\hat{\sigma}^2/2) > 1$. Para um $\hat{\sigma}^2$ grande, esse fator de ajuste poderá ser substancialmente maior que a unidade.

A previsão em (6.40) não é não-viesada, mas consistente. Não existem previsões não-viesadas de *y*, e em muitos casos (6.40) funciona bem. Porém, ela depende da normalidade do termo erro, *u*. No Capítulo 5, mostramos que o MQO possui propriedades desejáveis, mesmo quando *u* não for normalmente distribuído. Portanto, é vantajoso ter uma previsão que não dependa da normalidade. Se simplesmente assumirmos que *u* é independente das variáveis explicativas, teremos

$$E(y|\mathbf{x}) = \alpha_0 \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k), \qquad \textbf{(6.41)}$$

onde α_0 é o valor esperado de exp(*u*), que deve ser maior que a unidade.

Dada uma estimativa de $\hat{\alpha}_0$, podemos prever y como

$$\hat{y} = \hat{\alpha}_0 \exp(\hat{logy}), \tag{6.42}$$

que mais uma vez simplesmente requer que façamos a exponenciação do valor previsto do modelo log e que multipliquemos o resultado por $\hat{\alpha}_0$.

O resultado é que um estimador consistente de $\hat{\alpha}_0$ é facilmente obtido.

PREVISÃO DE *y* QUANDO A VARIÁVEL DEPENDENTE É LOG(*y*):

(i) Obtenha os valores estimados $\hat{logy}_i$ da regressão de $logy$ sobre $x_1,..., x_k$.
(ii) Para cada observação i, crie $\hat{m}_i = \exp(\hat{logy}_i)$.
(iii) Agora, faça a regressão de y sobre a variável única $\hat{m}$ *sem* um intercepto; isto é, faça uma regressão simples passando pela origem. O coeficiente de $\hat{m}$, o único coeficiente que existe, é a estimativa de α_0.

Quando $\hat{\alpha}_0$ for obtido, ele poderá ser usado juntamente com previsões de $logy$ para prever y. Os passos são os seguintes:

(i) Para valores dados de $x_1, x_2, ..., x_k$, obtenha $\hat{logy}$ de (6.39).
(ii) Obtenha a previsão de $\hat{y}$ de (6.42).

EXEMPLO 6.7

(Previsão de Salários de Diretores Executivos)

O modelo de interesse é

$$\log(salário) = \beta_0 + \beta_1 \log(vendas) + \beta_2 \log(valmerc) + \beta_3 permceo + u,$$

de forma que β_1 e β_2 são elasticidades e $100 \cdot \beta_3$ é uma semi-elasticidade. A equação estimada utilizando os dados contidos no arquivo CEOSAL2.RAW é

$$\widehat{lsalário} = 4{,}504 + 0{,}163\ lvendas + 0{,}109\ lvalmerc + 0{,}0117\ permceo$$
$$(0{,}257)\ \ (0{,}039)\qquad (0{,}050)\qquad (0{,}0053) \tag{6.43}$$
$$n = 177, R^2 = 0{,}318,$$

onde, para maior clareza, *lsalário* representa o log de *salário* e, de forma semelhante, *lvendas* e *lvalmerc* representam o log de *vendas* e *valmerc*. A seguir, obtemos $\hat{m}_i = \exp(\widehat{lsalário}_i)$ de cada observação na amostra. A regressão de *salário* sobre $\hat{m}$ (sem uma constante) produz $\hat{\alpha}_0 \approx 1{,}117$.

Podemos usar esse valor de $\hat{\alpha}_0$ juntamente com (6.43) para prever *salário* de qualquer valor de *vendas*, *valmerc* e *permceo*. Encontremos a previsão de *vendas* = 5.000 (que significa 5 bilhões de dólares, já que *vendas* está em milhões de dólares), *valmerc* = 10.000 (10 bilhões de dólares), e *permceo* = 10. Da equação (6.43), a previsão de *lsalário* é $4{,}504 + 0{,}163 \cdot \log(5.000) + 0{,}109 \cdot \log(10.000) + 0{,}0117(10) \approx 7{,}013$. O salário previsto é, portanto, $1{,}117 \cdot \exp(7{,}013) \approx 1.240{,}967$, ou 1.240.967 dólares. Se esquecermos de multiplicar por $\hat{\alpha}_0 = 1{,}117$, obteremos uma previsão de 1.110.983 dólares.

Podemos usar o método anterior de obter previsões para determinar o quanto o modelo com log(y) como variável dependente explica bem a variável y. Já temos indicadores para modelos quando y é a variável dependente: o R-quadrado e o R-quadrado ajustado. O objetivo é encontrar um bom indicador de grau de ajuste no modelo log(y) que possa ser comparado com um R-quadrado de um modelo no qual y é a variável dependente.

Existem várias maneiras de encontrarmos esse indicador, mas apresentamos um método que é fácil de ser implementado. Após computar a regressão de y sobre $\hat{m}$ passando pela origem no passo (iii), obtemos os valores estimados dessa regressão, $\hat{y}_i = \hat{\alpha}_0 m_i$. Em seguida, encontramos a correlação amostral entre $\hat{y}_i$ e o verdadeiro y_i na amostra. O quadrado dessa correlação amostral pode ser comparado com o R-quadrado que obtemos com o uso de y como a variável dependente em um modelo de regressão linear. Lembre-se de que o R-quadrado na equação estimada

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

é exatamente a correlação quadrada entre y_i e $\hat{y}_i$ (veja a Seção 3.2).

EXEMPLO 6.8

(Previsão de Salários de Diretores Executivos)

Após o passo (iii) no procedimento precedente, obtemos os valores estimados $\widehat{salário} = \hat{\alpha}_0 \hat{m}_i$. A correlação simples entre $salário_i$ e $\widehat{salário}$ na amostra é 0,493; o quadrado desse valor está por volta de 0,243. Esse é o nosso indicador de quanto da variação de *salário* é explicada pelo modelo log; não se trata do R-quadrado de (6.43), que é 0,318.

Suponha que estimemos um modelo com todas as variáveis em nível:

$$salário = \beta_0 + \beta_1 vendas + \beta_2 valmerc + \beta_3 permceo + u.$$

O R-quadrado obtido estimando esse modelo, usando as mesmas 177 observações, é 0,201. Assim, o modelo log explica mais da variação em *salário*, e assim é o nosso preferido com base no grau de ajuste. O modelo log também é o escolhido porque parece ser mais realista e seus parâmetros são mais fáceis de ser interpretados.

Neste capítulo, tratamos de alguns tópicos importantes sobre a análise de regressão múltipla.

A Seção 6.1 mostrou que uma mudança nas unidades de medida de uma variável independente altera o coeficiente do MQO da maneira esperada: se x_j for multiplicado por c, seu coeficiente será dividido por c. Se a variável dependente é multiplicada por c, todos os coeficientes de MQO são multiplicados por c. Nem a estatística t nem a F são alteradas pela mudança das unidades de medida de quaisquer variáveis.

Discutimos os coeficientes beta, que medem os efeitos das variáveis independentes sobre a variável dependente em unidades de desvios-padrão. Os coeficientes beta são obtidos de uma regressão

MQO padrão depois de as variáveis dependente e independentes terem sido transformadas em valores padronizados.

Como vimos em vários exemplos, a forma funcional logarítmica produz coeficientes com interpretações de efeitos percentuais. Comentamos sobre suas vantagens adicionais na Seção 6.2. Também vimos como computar o efeito percentual exato quando um coeficiente em um modelo log-nível é grande. Modelos com termos quadráticos consideram efeitos marginais decrescentes ou crescentes. Modelos com interações possibilitam que o efeito marginal de uma variável explicativa dependa do nível de outra variável explicativa.

Introduzimos o R-quadrado ajustado, $\bar{R}^2$, como uma alternativa ao R-quadrado habitual para medir o grau de ajuste. Enquanto o R^2 nunca pode cair quando outra variável é adicionada na regressão, o $\bar{R}^2$ penaliza o número de regressores e pode cair quando uma variável independente é adicionada. Isso faz do $\bar{R}^2$ o preferido para a opção entre modelos não-aninhados com diferentes quantidades de variáveis explicativas. Nem o R^2 nem o $\bar{R}^2$ podem ser usados para comparar modelos com variáveis dependentes diferentes. No entanto, é bastante fácil obter indicadores de graus de ajuste para optarmos entre y e $\log(y)$ como a variável dependente, como mostrado na Seção 6.4.

Na Seção 6.3 discutimos o problema de certa forma sutil de dependermos demasiadamente do R^2 ou do $\bar{R}^2$ para chegarmos a um modelo final: é possível controlarmos grandes quantidades de fatores em um modelo de regressão. Por essa razão, é importante pensar à frente sobre a especificação de modelos, particularmente sobre a natureza *ceteris paribus* da equação de regressão múltipla. Variáveis explicativas que afetem y e sejam não-correlacionadas com todas as outras variáveis explicativas podem ser usadas para reduzir a variância do erro sem induzir multicolinearidade.

Na Seção 6.4 demonstramos como obter um intervalo de confiança de uma previsão feita de uma linha de regressão MQO. Também mostramos como um intervalo de confiança pode ser construído para um valor futuro e desconhecido de y.

Ocasionalmente, queremos prever y quando $\log(y)$ é usado como a variável dependente em um modelo de regressão. A Seção 6.4 explica esse método simples. Finalmente, algumas vezes estamos interessados em conhecer o sinal e a magnitude dos resíduos de observações específicas. A análise de resíduos pode ser usada para determinarmos se elementos específicos da amostra possuem valores previstos que estejam bem acima, ou bem abaixo, dos verdadeiros resultados.

6.1 A seguinte equação foi estimada utilizando os dados contidos no arquivo CEOSAL1.RAW:

$$\widehat{\log(salário)} = \underset{(0,324)}{4,322} + \underset{(0,33)}{0,276}\log(vendas) + \underset{(0,129)}{0,0215}\,rma - \underset{(0,00026)}{0,00008}\,rma^2$$

$$n = 209,\ R^2 = 0,282.$$

Esta equação permite que *rma* tenha um efeito decrescente sobre log(*salário*). Essa generalidade é necessária? Justifique.

6.2 Sejam $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ as estimativas MQO da regressão de y_i sobre $x_{i1}, ..., x_{ik}$, $i = 1, 2, ..., n$. Para constantes diferentes de zero $c_1, ..., c_k$, argumente que o intercepto e as inclinações MQO da regressão de $c_0 y_i$ sobre $c_1 x_{i1}, ..., c_k x_{ik}$, $i = 1, 2, ..., n$ são dados por $\tilde{\beta}_0 = c_0 \hat{\beta}_0$, $\tilde{\beta}_1 = (c_0/c_1)\hat{\beta}_1, ..., \tilde{\beta}_k = (c_0/c_k)$

$\hat{\beta}_k$. (*Sugestão*: Use o fato de que $\hat{\beta}_j$ soluciona as condições de primeira ordem em (3.13), e que $\tilde{\beta}_j$ deve solucionar as condições de primeira ordem envolvendo as variáveis dependente e independentes redimensionadas.)

6.3 Utilizando os dados contidos no arquivo RDCHEM.RAW, a seguinte equação foi obtida por MQO:

$$\widehat{pdintens} = \underset{(0,429)}{2,613} + \underset{(0,00014)}{0,00030}\, vendas - \underset{(0,0000000037)}{0,0000000070}\, vendas^2$$

$$n = 32,\ R^2 = 0,1484.$$

(i) Em que ponto o efeito marginal de *vendas* sobre *pdintens* se torna negativo?

(ii) Você manteria o termo quadrático no modelo? Explique.

(iii) Defina *vendasbil* como vendas expressas em bilhões de dólares: *vendasbil* = *vendas*/1.000. Reescreva a equação com *vendasbil* e $vendasbil^2$ como as variáveis independentes. Certifique-se de descrever os erros-padrão e o R-quadrado. [*Sugestão*: Observe que $vendasbil^2 = vendas^2/(1.000)^2$.]

(iv) Com o propósito de descrever os resultados, qual equação você prefere?

6.4 O seguinte modelo permite que o retorno da educação dependa da educação total dos pais, chamada *edupais*:

$$\log(salário) = \beta_0 + \beta_1 educ + \beta_2\, educ \cdot edupais + \beta_3 exper + \beta_4 perm + u.$$

(i) Mostre que, em forma decimal, o retorno de mais um ano de educação nesse modelo é

$$\Delta\log(salário)/\Delta educ = \beta_1 + \beta_2 edupais.$$

Que sinal você espera para β_2? Por quê?

(ii) Utilizando os dados contidos no arquivo WAGE2.RAW, a equação estimada é

$$\widehat{\log(salário)} = \underset{(0,13)}{5,65} + \underset{(0,010)}{0,047}\, educ + \underset{(0,00021)}{0,00078}\, educ \cdot edupais +$$

$$\underset{(0,004)}{0,019}\, exper + \underset{(0,003)}{0,010}\, perm$$

$$n = 722,\ R^2 = 0,169.$$

(Somente 722 observações contêm todas as informações sobre a educação dos pais.) Interprete o coeficiente do termo de interação. Pode ser interessante escolher dois valores específicos para *edupais*, por exemplo, *edupais* = 32 se ambos tiverem educação superior, ou *edupais* = 24 se ambos tiverem educação de nível médio — e comparar o retorno estimado de *educ*.

(iii) Quando *edupais* é adicionada como uma variável separada na equação, obtemos:

$$\widehat{\log(salário)} = 4{,}94 + 0{,}097\,educ + 0{,}033\,edupais + 0{,}0016\,educ\cdot edupais$$
$$(0{,}38)\ (0{,}027)\qquad (0{,}017)\qquad (0{,}0012)$$
$$+ 0{,}020\,exper + 0{,}010\,perm$$
$$(0{,}004)\qquad (0{,}003)$$
$$n = 722,\ R^2 = 0{,}174.$$

O retorno da educação agora depende positivamente da educação dos pais? Teste a hipótese nula de que o retorno da educação não depende da educação dos pais.

6.5 No Exemplo 4.2, no qual a percentagem de alunos aprovados em um exame de matemática do 10º ano (*mate10*) é a variável dependente, faz sentido incluir *cien11* — a percentagem de alunos do 11º ano aprovados em um exame de ciências — como uma variável explicativa adicional?

6.6 Quando $taxafreq^2$ e $tac\cdot taxafreq$ são adicionadas à equação (6.19), o R-quadrado passa a ser 0,232. Esses termos adicionais são conjuntamente significantes no nível de 10%? Você os incluiria no modelo?

6.7 As três seguintes equações foram estimadas utilizando-se as 1,534 observações contidas no arquivo 401K.RAW.

$$\widehat{taxap} = 80{,}29 + 5{,}44\,taxcomp + 0{,}269\,idade - 0{,}00013\,totemp$$
$$(0{,}78)\ (0{,}52)\qquad (0{,}045)\qquad (0{,}00004)$$
$$R^2 = 0{,}100,\ \overline{R}^2 = 0{,}098.$$

$$\widehat{taxap} = 97{,}32 + 5{,}02\,taxcomp + 0{,}314\,idade - 2{,}66\log(totemp)$$
$$(1{,}95)\ (0{,}51)\qquad (0{,}044)\qquad (0{,}28)$$
$$R^2 = 0{,}144,\ \overline{R}^2 = 0{,}142.$$

$$\widehat{taxap} = 80{,}62 + 5{,}34\,taxcomp + 0{,}290\,idade - 0{,}00043\,totemp$$
$$(0{,}78)\ (0{,}52)\qquad (0{,}045)\qquad (0{,}00009)$$
$$+ 0{,}0000000039\,totemp^2$$
$$(0{,}0000000010)$$
$$R^2 = 0{,}108,\ \overline{R}^2 = 0{,}106.$$

Qual desses três modelos você prefere? Por quê?

Análise de Regressão Múltipla com Informações Qualitativas: Variáveis Binárias (ou *Dummy*)

os capítulos anteriores, as variáveis dependentes e independentes em nossos modelos de regressão múltipla tinham significado *quantitativo*. Alguns exemplos incluíam taxas de salário por hora, anos de escolaridade, nota média em curso superior, quantidade da poluição do ar, níveis de vendas de empresas e número de detenções. Em cada caso, a magnitude da variável carrega informações valiosas. No trabalho empírico também devemos incorporar fatores *qualitativos* nos modelos de regressão. O sexo ou a raça de um indivíduo, o ramo de atividade de uma empresa (fabricante, varejista etc.) e a região onde uma cidade está localizada (sul, norte, oeste etc.) são todos considerados fatores qualitativos.

A maior parte deste capítulo é dedicada a variáveis *independentes* qualitativas. Após discutirmos as maneiras apropriadas de descrever informações qualitativas na Seção 7.1, mostraremos como variáveis explicativas qualitativas podem ser facilmente incorporadas em modelos de regressão múltipla nas Seções 7.2, 7.3 e 7.4. Essas seções tratam de quase todos os modos conhecidos nos quais as variáveis independentes qualitativas são usadas na análise de regressão de corte transversal.

Na Seção 7.5 examinaremos uma variável dependente binária, que é um tipo especial de variável dependente qualitativa. O modelo de regressão múltipla tem uma interpretação bastante interessante neste caso e é chamado de modelo de probabilidade linear. Embora muito criticado por alguns econometristas, a simplicidade do modelo de probabilidade linear faz dele uma ferramenta útil em muitos contextos empíricos. Também descreveremos suas falhas na Seção 7.5, embora elas sejam freqüentemente secundárias no trabalho empírico.

7.1 A DESCRIÇÃO DAS INFORMAÇÕES QUALITATIVAS

Fatores qualitativos freqüentemente aparecem na forma de informação binária: uma pessoa é do sexo feminino ou masculino; alguém possui ou não um computador pessoal; uma firma oferece ou não certo tipo de plano de pensão a seus empregados; um estado adota ou não a pena capital. Em todos esses exemplos, a informação relevante pode ser capturada pela definição de uma **variável binária** ou uma variável zero-um. Em econometria, as variáveis binárias são em geral chamadas **variáveis *dummy***, embora esse nome não seja muito descritivo.

Ao definirmos uma variável *dummy*, precisamos decidir a qual evento será atribuído o valor um e a qual será atribuído o valor zero. Por exemplo, em um estudo sobre a determinação do salário individual, podemos definir *feminino* como a variável binária que assumirá o valor um quando a pessoa for mulher e zero, quando homem. Neste caso, o nome indica o evento cujo valor é um. A mesma informação é transmitida se definirmos que masculino será um, se a pessoa for homem, e zero, se mulher.

Qualquer uma dessas formas é melhor que usarmos *gênero*, porque esse nome não deixa claro quando a variável *dummy* é um: gênero = 1 corresponde a homem ou a mulher? A maneira pela qual denominamos nossas variáveis não tem importância para obtermos os resultados da regressão, mas sempre ajuda a escolhermos nomes que deixem claras as equações e as explicações.

Suponha que, em um estudo que compara resultados de eleições entre candidatos democratas e republicanos, você queira indicar o partido de cada candidato. Um nome como *partido* será uma boa escolha para uma variável binária neste caso? Qual seria um nome melhor?

Suponha que no exemplo do salário tenhamos escolhido o nome *feminino* para indicar o sexo. Além disso, definimos uma variável binária *casado* como igual a um se a pessoa for casada, e zero, caso contrário. A Tabela 7.1 fornece uma listagem parcial de um possível conjunto de dados sobre salários. Vemos que a Pessoa 1 é do sexo feminino e não é casada, a Pessoa 2 é do sexo feminino e é casada, a Pessoa 3 é do sexo masculino e não é casada, e assim por diante.

Por que usamos os valores zero e um para descrever informações qualitativas? Em certo sentido, esses valores são arbitrários: quaisquer dois valores diferentes serviriam. O benefício real de capturar informação qualitativa usando variáveis zero-um é que elas levam a modelos de regressão nos quais os parâmetros têm interpretações bastante naturais, como veremos agora.

Tabela 7.1

Uma Listagem Parcial dos Dados do Arquivo WAGE1.RAW

pessoa	*salárioh*	*educ*	*exper*	*feminino*	*casado*
1	3,10	11	2	1	0
2	3,24	12	22	1	1
3	3,00	11	2	0	0
4	6,00	8	44	0	1
5	5,30	12	7	0	1
. . .	. . .	. . .	. . .	. . .	. . .
525	11,56	16	5	0	1
526	3,50	14	5	1	0

7.2 UMA ÚNICA VARIÁVEL *DUMMY* INDEPENDENTE

Como incorporamos informações binárias em modelos de regressão? No caso mais simples, com somente uma variável *dummy* explicativa, simplesmente adicionamos a variável à equação como uma variável independente. Por exemplo, considere o seguinte modelo simples de determinação de salários por hora:

$$salárioh = \beta_0 + \delta_0 feminino + \beta_1 educ + u. \quad \textbf{(7.1)}$$

Usamos δ_0 como o parâmetro da variável *feminino* de maneira a ressaltar a interpretação dos parâmetros que multiplicam variáveis *dummy*; mais adiante, usaremos a notação que for mais conveniente.

No modelo (7.1), somente dois fatores observados afetam os salários: gênero e educação. Como *feminino* = 1 quando a pessoa é mulher e *feminino* = 0 quando a pessoa é homem, o parâmetro δ_0 tem a seguinte interpretação: δ_0 é a diferença no salário por hora entre mulheres e homens, *dado* o mesmo grau de educação (e o mesmo termo erro *u*). Assim, o coeficiente δ_0 determina se existe discriminação contra as mulheres: se $\delta_0 < 0$, então, para o mesmo nível dos outros fatores, as mulheres ganham menos que os homens, em média.

Em termos de expectativas, se assumirmos a hipótese de média condicional zero $E(u|feminino,educ) = 0$, então

$$\delta_0 = E(salárioh|feminino = 1,educ) - E(salárioh|feminino = 0, educ).$$

Como *feminino* = 1 corresponde a mulheres e *feminino* = 0 corresponde a homens, podemos escrever essa expressão de forma mais simples:

$$\delta_0 = E(salárioh|feminino,educ) - E(salárioh|masculino,educ). \quad \textbf{(7.2)}$$

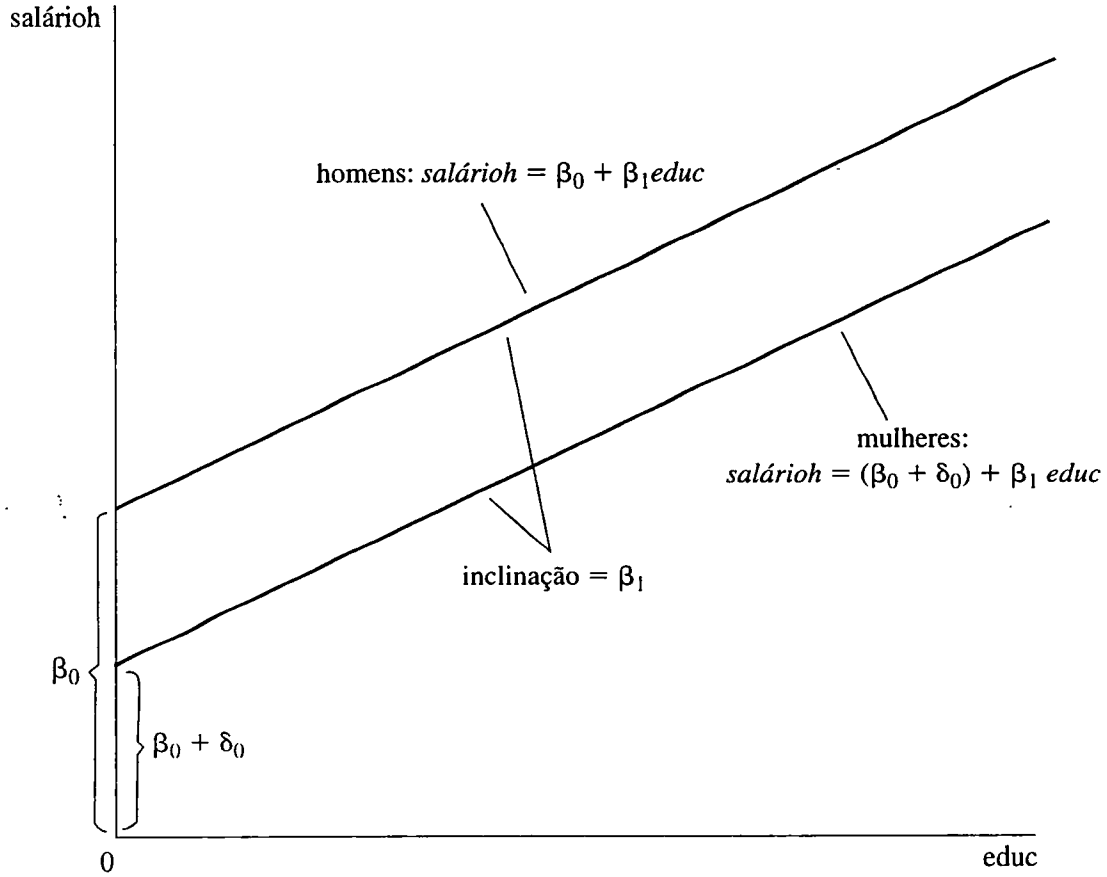
O importante aqui é que o nível de educação é o mesmo em ambas as expectativas; a diferença, δ_0, deve-se somente ao gênero.

A situação pode ser descrita graficamente como um **deslocamento de intercepto** entre as linhas que representam homens e mulheres. Na Figura 7.1, o caso $\delta_0 < 0$ é mostrado, de modo que os homens ganham um montante fixo por hora a mais que as mulheres. A diferença não depende do nível de educação, e isso explica a razão de os perfis salário-educação das mulheres e dos homens serem paralelos.

Neste ponto, você pode estar se perguntando por que não incluímos, também, em (7.1) uma variável *dummy*, digamos, *masculino*, que seria um para homens e zero para mulheres. A razão é que isso seria redundante. Na equação (7.1), o intercepto para homens é β_0, enquanto o intercepto para mulheres é $\beta_0 + \delta_0$. Como existem apenas dois grupos, precisamos de apenas dois interceptos diferentes. Isso significa que, além de β_0, precisamos usar somente *uma* variável *dummy*; decidimos incluir a variável *dummy* para mulheres. O uso de duas variáveis *dummy* introduziria colinearidade perfeita, porque *feminino* + *masculino* = 1, o que significa que *masculino* é uma função linear perfeita de *feminino*. A inclusão de variáveis *dummy* para ambos os sexos é o exemplo mais simples da chamada **armadilha da variável *dummy***, que surge quando um grande número de variáveis *dummy* descreve determinado número de grupos. Discutiremos esse assunto mais adiante.

Figura 7.1

Gráfico de $salárioh = \beta_0 + \delta_0 feminino + \beta_1 educ$ para $\delta_0 < 0$.



Na equação (7.1) acima, escolhemos homens para ser o **grupo base** ou o **grupo de referência**, isto é, o grupo contra o qual as comparações são feitas. Esta é a razão pela qual β_0 é o intercepto para os homens, e δ_0 é a *diferença* dos interceptos entre mulheres e homens. Poderíamos ter escolhido as mulheres como o grupo base, escrevendo o modelo como

$$salárioh = \alpha_0 + \gamma_0 masculino + \beta_1 educ + u,$$

onde o intercepto para mulheres é α_0 e o intercepto para homens é $\alpha_0 + \gamma_0$; isso implica que $\alpha_0 = \beta_0 + \delta_0$ e $\alpha_0 + \gamma_0 = \beta_0$. Em qualquer aplicação, não importa como escolhemos o grupo base, mas é importante estar atento para qual é o grupo base.

Alguns pesquisadores preferem eliminar o intercepto global do modelo e incluir variáveis *dummy* para cada grupo. A equação então seria $salárioh = \beta_0 masculino + \alpha_0 feminino + \beta_1 educ + u$, onde o intercepto para os homens é β_0 e o intercepto para as mulheres é α_0. Não existe armadilha da variável *dummy* neste caso, porque não temos um intercepto global. Porém, essa formulação tem pouco a oferecer, já que é mais difícil verificar diferenças nos interceptos, e não existe uma maneira consensual de computar o R-quadrado em regressões sem intercepto. Portanto, sempre incluiremos um intercepto global para o grupo base.

Nada mais muda muito quando mais variáveis explicativas estão envolvidas. Considerando os homens como o grupo base, um modelo que controla a experiência e a permanência, além da educação é

$$salárioh = \beta_0 + \delta_0 feminino + \beta_1 educ + \beta_2 exper + \beta_3 perm + u. \quad (7.3)$$

Se *educ*, *exper* e *perm* forem todas características relevantes da produtividade, a hipótese nula de não-existência de diferença entre homens e mulheres será H_0: $\delta_0 = 0$. A hipótese alternativa de que existe discriminação contra as mulheres será H_1: $\delta_0 < 0$.

Como podemos efetivamente testar a discriminação salarial? A resposta é simples: simplesmente estimamos o modelo por MQO, *exatamente* como antes, e usamos a estatística *t* habitual. Nada muda na mecânica do MQO ou na teoria estatística quando algumas das variáveis independentes são definidas como variáveis *dummy*. A única diferença em relação ao que vínhamos fazendo até agora é a interpretação do coeficiente da variável *dummy*.

EXEMPLO 7.1

(Equação dos Salários por Hora)

Utilizando os dados contidos no arquivo WAGE1.RAW, estimamos o modelo (7.3). Por enquanto, usamos *salárioh*, em vez de log(*salárioh*), como a variável dependente:

$$\begin{aligned} salárioh = &-1{,}57 - 1{,}81\, feminino + 0{,}572\, educ \\ &\quad (0{,}72)\ (0{,}26) \qquad\quad (0{,}049) \\ &+ 0{,}025\, exper + 0{,}141\, perm \\ &\quad (0{,}012) \qquad (0{,}021) \\ &n = 526,\ R^2 = 0{,}364. \end{aligned} \quad (7.4)$$

O intercepto negativo — o intercepto para os homens, neste caso — não é muito significativo, já que ninguém na amostra tem anos de *educ*, *exper* e *perm* próximos de zero. O coeficiente de *feminino* é interessante, porque ele registra a diferença média no salário por hora entre uma mulher e um homem, dados os *mesmos* níveis de *educ*, *exper* e *perm*. Se compararmos uma mulher e um homem com os mesmos níveis de educação, experiência e permanência, a mulher ganha, em média, 1,81 dólares por hora a menos que o homem. (Não se esqueça de que estamos tratando de salários de 1976.)

É importante lembrarmos que, como fizemos uma regressão múltipla e controlamos *educ*, *exper* e *perm*, o diferencial de 1,81 dólares no salário não pode ser explicado por diferentes níveis médios de educação, experiência ou permanência entre homens e mulheres. Podemos concluir que o diferencial de 1,81 dólares é devido ao gênero ou a fatores associados ao gênero que não tenhamos controlado na regressão.

É esclarecedor comparar o coeficiente de *feminino* na equação (7.4) com a estimativa que obtemos quando todas as outras variáveis explicativas são eliminadas da equação:

EXEMPLO 7.1 (continuação)

$$\begin{gathered} salárioh = 7{,}10 - 2{,}51\, feminino \\ (0{,}21)\quad (0{,}30) \\ n = 526,\ R^2 = 0{,}116 \end{gathered} \tag{7.5}$$

Os coeficientes em (7.5) têm uma interpretação simples. O intercepto é o salário-hora médio dos homens na amostra (*feminino* = 0), de modo que os homens ganham, em média, 7,10 dólares por hora. O coeficiente de *feminino* é a diferença no salário médio entre homens e mulheres. Assim, o salário médio das mulheres, na amostra, é 7,10 − 2,51 = 4,59, ou 4,59 dólares por hora. (A propósito, existem 274 homens e 252 mulheres na amostra.)

A equação (7.5) apresenta um modo simples de realizar um *teste de comparação de médias* entre os dois grupos, que neste caso são homens e mulheres. A diferença estimada −2,51 tem uma estatística *t* de −8,37, que é, estatisticamente, bastante significante (e, claro, 2,51 dólares também é grande economicamente). Geralmente, a regressão simples sobre uma constante e uma variável *dummy* é uma maneira bastante objetiva de comparar as médias de dois grupos. Para que o teste *t* habitual seja válido, temos que assumir a manutenção da hipótese de homoscedasticidade, o que significa que a variância populacional dos salários dos homens é a mesma dos salários das mulheres.

O diferencial salarial estimado entre homens e mulheres é maior em (7.5) que na equação (7.4), porque (7.5) não controla as diferenças em educação, experiência e permanência, e esses fatores são mais baixos, em média, para as mulheres do que para os homens nessa amostra. A equação (7.4) fornece uma estimativa mais confiável da discrepância salarial *ceteris paribus* entre os sexos; ela ainda indica um diferencial bastante grande.

Em muitos casos, variáveis *dummy* independentes refletem escolhas de indivíduos ou de outras unidades econômicas (em oposição a algo predeterminado, como gênero). Em tais situações, a questão da causalidade é novamente crucial. No exemplo seguinte, gostaríamos de saber se o fato de possuir um computador pessoal *causa* uma nota média mais elevada em curso superior.

EXEMPLO 7.2

(Efeitos de se Possuir Computadores na Avaliação em Cursos Superiores)

Para determinar os efeitos de se possuir um computador na nota média em curso superior, estimamos o modelo

$$nmgrad = \beta_0 + \delta_0 PC + \beta_1 nmem + \beta_2 tac + u,$$

onde a variável *dummy PC* é igual a um se um aluno possui um computador pessoal e zero caso contrário. Existem várias razões pelas quais a propriedade de um computador pessoal pode ter um efeito sobre *nmgrad*. O trabalho de um aluno pode ser de melhor qualidade se feito em um computador e ele pode ganhar tempo por não ter que ficar esperando sua vez em um laboratório de informática. É claro que o aluno pode estar mais inclinado a brincar com jogos ou a navegar na Internet se ele, ou ela, possuir seu próprio PC,

EXEMPLO 7.2 (continuação)

de modo que não é óbvio que δ_0 será positivo. As variáveis *nmem* (nota média do ensino médio) e *tac* (nota do teste de avaliação para ingresso em curso superior) são usadas como controles: pode ser que alunos mais fortes, medidos pelas nota média do ensino médio e nota do teste de avaliação, provavelmente possuam computadores. Controlamos esses fatores porque gostaríamos de saber o efeito médio sobre *nmgrad* se escolhermos um aluno aleatoriamente e dermos a ele um computador pessoal.

Utilizando os dados contidos no arquivo GPA1.RAW, obtemos

$$\widehat{nmgrad} = 1{,}26 + 0{,}157\,PC + 0{,}447\,nmem + 0{,}0087\,tac \quad \textbf{(7.6)}$$
$$(0{,}33)\ (0{,}057)\qquad (0{,}094)\qquad (0{,}0105)$$
$$n = 141,\ R^2 = 0{,}219.$$

Esta equação sugere para um aluno que possua um computador pessoal uma nota média prevista em torno de 0,16 pontos mais alta quando comparada com a de um aluno que não possui um PC (lembre-se, tanto *nmgrad* como *nmem* estão em uma escala de quatro pontos). O efeito também é, estatisticamente, bastante significante, com $t_{PC} = 0{,}157/0{,}057 \approx 2{,}75$.

O que acontece se eliminarmos *nmem* e *tac* da equação? Claramente, a eliminação da última variável deve ter um efeito muito pequeno, já que seu coeficiente e a respectiva estatística *t* são muito pequenos. Mas *nmem* é bastante significante, e assim sua eliminação pode afetar a estimativa de β_{PC}. A regressão de *nmgrad* sobre *PC* produz a estimativa do coeficiente de *PC* igual a, aproximadamente, 0,170, com um erro-padrão de 0,063; neste caso, $\hat{\beta}_{PC}$ e sua estatística *t* não mudam muito.

Nos exercícios no final do capítulo será solicitado o controle de outros fatores na equação para verificar se o efeito de se possuir um computador desaparece, ou se pelo menos se torna significativamente menor.

Cada um dos exemplos anteriores pode ser entendido como relevante para a **análise de políticas públicas**. No primeiro exemplo estávamos interessados na discriminação sexual na força de trabalho. No segundo exemplo estávamos preocupados com o efeito de se possuir um computador sobre o desempenho no curso superior. Um caso especial de análise de políticas públicas é a **avaliação de programas**, na qual gostaríamos de saber o efeito de programas econômicos ou sociais sobre os indivíduos, empresas, vizinhança, cidades etc.

No caso mais simples existem dois grupos de objetos de estudo. O **grupo de controle** não participa do programa. O **grupo experimental** ou **grupo de tratamento** faz parte do programa. Esses nomes provêm da literatura das ciências experimentais, não devem ser interpretados literalmente. Exceto em casos raros, a escolha dos grupos de controle e de tratamento não é feita aleatoriamente. Porém, em alguns casos, a análise de regressão múltipla pode ser usada para controlar um número suficiente de outros fatores para estimar o efeito causal do programa.

EXEMPLO 7.3

(Efeitos da Concessão de Subsídios sobre as Horas de Treinamento)

Utilizando os dados de 1988 das indústrias de Michigan contidas no arquivo JTRAIN.RAW, obteremos a seguinte equação estimada:

EXEMPLO 7.3 (continuação)

$$\begin{aligned} \widehat{hrsemp} &= 46{,}67 + 26{,}25\ subs - 0{,}98\ \log(vendas) \\ &\quad (43{,}41)\quad (5{,}59)\qquad (3{,}54) \\ &\quad - 6{,}07\ \log(empreg) \\ &\quad (3{,}88) \\ n &= 105,\ R^2 = 0{,}237. \end{aligned} \tag{7.7}$$

A variável dependente é horas de treinamento por empregado, ao nível da empresa. A variável *subs* é uma variável *dummy* igual a um se a firma recebeu um subsídio para treinamento em 1988 e zero, caso contrário. As variáveis *vendas* e *empreg* representam as vendas anuais e o número de empregados, respectivamente. Não podemos usar *hrsemp* na forma logarítmica porque ela tem valor zero para 20 das 105 empresas usadas na regressão.

A variável *subs* é estatisticamente bastante significante, com $t_{subs} = 4{,}70$. Controlando vendas e emprego, as empresas que receberam subsídios treinaram cada um de seus empregados, em média, 26,25 horas mais que as outras. Como o número médio de horas de treinamento por empregado na amostra está em torno de 17, com um valor máximo de 164, *subs* tem um grande efeito sobre o treinamento, como o esperado.

O coeficiente de log(*vendas*) é pequeno e não significante. O coeficiente de log(*empreg*) significa que, se a empresa for 10% maior, ela treinará seus empregados cerca de 0,61 horas menos. Sua estatística *t* é −1,56, que é somente marginalmente significante, em termos estatísticos.

Assim como em relação a qualquer outra variável independente, devemos perguntar se o efeito mensurado de uma variável qualitativa é causal. Na equação (7.7), a diferença de treinamento entre as empresas que recebem subsídios e as que não recebem se deve aos subsídios, ou o recebimento de um subsídio é simplesmente um indicador de alguma outra coisa? É possível que as empresas que recebem subsídios tenham, em média, treinado seus empregados de modo mais regular, mesmo sem os subsídios. Nada nesta análise nos informa se estimamos um efeito causal; precisamos saber como foram determinadas as empresas que receberiam subsídios. Somente podemos esperar que tenhamos controlado tantos fatores quanto possível de modo que possam estar relacionados à questão de a empresa ter recebido subsídios e com os seus níveis de treinamento.

Retornaremos à análise de políticas públicas com variáveis *dummy* na Seção 7.6, como também em outros capítulos.

A Interpretação dos Coeficientes de Variáveis *Dummy* Explicativas quando a Variável Dependente É Expressa como log(*y*)

Uma especificação comum em trabalhos aplicados tem a variável dependente aparecendo na forma logarítmica, com uma ou mais variáveis *dummy* aparecendo como variáveis independentes. Como interpretamos os coeficientes das variáveis *dummy* neste caso? Não surpreendentemente, os coeficientes têm uma interpretação *percentual*.

EXEMPLO 7.4

(Regressão dos Preços de Imóveis)

Utilizando os dados contidos no arquivo HPRICE1.RAW, obtemos a equação

$$\log(\widehat{preço}) = 5,56 + 0,168\log(tamterr) + 0,707\log(arquad)$$
$$(0,65)\ (0,038)\qquad (0,093)$$
$$+ 0,027\,qtdorm + 0,054\,colonial \qquad (7.8)$$
$$(0,029)\qquad (0,045)$$
$$n = 88, R^2 = 0,649.$$

Todas as variáveis são auto-explicativas, exceto *colonial*, que é uma variável binária igual a um se o imóvel tiver estilo colonial. O que significa o coeficiente de *colonial*? Para níveis dados de *tamterr*, *arquad* e *qtdorm*, a diferença em log($\widehat{preço}$) entre um imóvel de estilo colonial e outro de outro estilo é 0,054. Isso significa prever que um imóvel de estilo colonial seja vendido por cerca de 5,4% a mais, mantendo-se todos os outros fatores fixos.

Este exemplo mostra que, quando log(y) é a variável dependente em um modelo, o coeficiente de uma variável *dummy*, quando multiplicado por 100, é interpretado como a diferença percentual em y, mantendo fixos todos os outros fatores. Quando o coeficiente de uma variável *dummy* sugere uma grande mudança proporcional em y, a diferença percentual exata pode ser obtida exatamente como no cálculo da semi-elasticidade na Seção 6.2.

EXEMPLO 7.5

(Equação Log do Salário-Hora)

Reestimemos a equação salarial do exemplo 7.1, usando log(*salárioh*) como a variável dependente e adicionando termos quadráticos em *exper* e *perm*:

$$\widehat{\log(salárioh)} = 0,417 - 0,297\,feminino + 0,080\,educ + 0,029\,exper$$
$$(0,099)\ (0,036)\qquad (0,007)\qquad (0,005)$$
$$- 0,00058\,exper^2 + 0,032\,perm - 0,00059\,perm^2 \qquad (7.9)$$
$$(0,00010)\qquad (0,007)\qquad (0,00023)$$
$$n = 526, R^2 = 0,441.$$

Usando a mesma aproximação do Exemplo 7.4, o coeficiente de *feminino* implica que, para os mesmos níveis de *educ*, *exper* e *perm*, as mulheres ganham cerca de 100(0,297) = 29,7% a menos que os homens. Podemos fazer melhor que isso ao computarmos a diferença percentual exata nos salários previstos. O que

EXEMPLO 7.5 (continuação)

queremos é a diferença proporcional nos salários entre mulheres e homens, mantendo fixos todos os outros fatores: $(\widehat{salário}_M - \widehat{salário}_H)/\widehat{salário}_H$. O que temos a partir de (7.9) é

$$\log(\widehat{salário}_M) - \log(\widehat{salário}_H) = 0{,}297.$$

Fazendo a exponenciação e a subtração temos

$$(\widehat{salário}_M - \widehat{salário}_H)/\widehat{salário}_H = \exp(-0{,}297) - 1 \approx -0{,}257.$$

Esta estimativa mais exata implica que o salário de uma mulher é, em média, 25,7% menor que o salário de um homem nas mesmas condições.

Se tivéssemos feito a mesma correção no Exemplo 7.4, teríamos obtido $\exp(0{,}054) - 1 \approx 0{,}0555$, ou cerca de 5,6%. A correção tem um efeito menor no Exemplo 7.4 do que no exemplo salarial, porque a magnitude do coeficiente da variável *dummy* é muito menor em (7.8) do que em (7.9).

De forma geral, se $\hat{\beta}_1$ for o coeficiente de uma variável *dummy*, digamos, x_1, quando $\log(y)$ é a variável dependente, a diferença percentual exata em y previsto quando $x_1 = 1$ *versus* quando $x_1 = 0$ é

$$100 \cdot [\exp(\hat{\beta}_1) - 1]. \qquad (7.10)$$

O coeficiente $\hat{\beta}_1$ estimado pode ser positivo ou negativo, e é importante preservar seu sinal ao computar (7.10).

7.3 O USO DE VARIÁVEIS *DUMMY* PARA CATEGORIAS MÚLTIPLAS

Podemos usar diversas variáveis *dummy* independentes na mesma equação. Por exemplo, poderíamos adicionar a variável *dummy casado* na equação (7.9). O coeficiente de *casado* fornece o diferencial proporcional (aproximado) nos salários entre aqueles que são, ou não, casados, mantendo fixos gênero, *educ*, *exper* e *perm*. Quando estimamos esse modelo, o coeficiente de *casado* (com o erro-padrão entre parênteses) é 0,053 (0,041), e o coeficiente de *feminino* passa a ser −0,290 (0,036). Assim, o "prêmio" por ser casado é estimado em torno de 5,3%, mas não é estatisticamente diferente de zero ($t = 1{,}29$). Uma limitação importante deste modelo é que o prêmio por ser casado é assumido como o mesmo para homens e mulheres; isso é relaxado no exemplo a seguir.

EXEMPLO 7.6

(Equação do Log do Salário-Hora)

Estimemos um modelo que considere diferenças salariais entre quatro grupos: homens casados, mulheres casadas, homens solteiros e mulheres solteiras. Para fazermos isso temos que selecionar um grupo base; escolhemos homens solteiros. Então, devemos definir as variáveis *dummy* para cada um dos demais grupos. Vamos chamá-los *hcasados*, *mcasadas* e *msolteiras*. Colocando essas três variáveis na equação (7.9) (e, claro, eliminando *feminino*, já que agora ela é redundante) produz

EXEMPLO 7.6 (continuação)

$$\begin{aligned}\log(\widehat{salário}) = {} & 0{,}321 + 0{,}213\,hcasados - 0{,}198\,mcasadas \\ & \quad (0{,}100)\ (0{,}055) \qquad\quad (0{,}058) \\ & - 0{,}110\,msolteiras + 0{,}079\,educ + 0{,}027\,exper - 0{,}00054\,exper^2 \\ & \quad (0{,}056) \qquad\quad (0{,}007) \qquad (0{,}005) \qquad\quad (0{,}00011) \\ & + 0{,}029\,perm - 0{,}00053\,perm^2 \\ & \quad (0{,}007) \qquad (0{,}00023) \\ & n = 526,\ R^2 = 0{,}461.\end{aligned} \tag{7.11}$$

Todos os coeficientes, exceto o de *msolteiras*, têm estatísticas *t* bem acima de dois, em valores absolutos. A estatística *t* de *msolteiras* está em torno de $-1{,}96$, que é significante apenas ao nível de 5% contra uma alternativa bilateral.

Para interpretar os coeficientes das variáveis *dummy*, devemos nos lembrar de que o grupo base é o de homens solteiros. Assim, as estimativas das três variáveis *dummy* medem a diferença proporcional nos salários *relativamente* aos homens solteiros. Por exemplo, estima-se que os homens casados ganhem cerca de 21,3% mais que os homens solteiros, mantendo fixas educação, experiência e permanência. [A estimativa mais precisa a partir de (7.10) está em torno de 23,7%.] Uma mulher casada, no entanto, deve ganhar 19,8% menos que um homem solteiro com os mesmos níveis das outras variáveis.

Como o grupo base é representado pelo intercepto na equação (7.11), incluímos variáveis *dummy* para apenas três dos quatro grupos. Se tivéssemos incluído uma variável *dummy* para homens solteiros em (7.11) cairíamos na armadilha da variável *dummy*, por termos introduzido colinearidade perfeita. Alguns programas de regressão corrigem automaticamente esse engano, enquanto outros apenas informam a existência de colinearidade perfeita. É melhor especificar cuidadosamente as variáveis *dummy*, pois isso nos forçará a interpretar apropriadamente o modelo final.

Embora os homens solteiros sejam o grupo base em (7.11), podemos usar essa equação para obter a diferença estimada entre dois grupos quaisquer. Como o intercepto global é comum a todos os grupos, podemos ignorá-lo quando procuramos diferenças. Assim, a diferença proporcional estimada entre as mulheres solteiras e as casadas é $-0{,}110 - (-0{,}198) = 0{,}088$, o que significa que as mulheres solteiras ganham cerca de 8,8% mais que as mulheres casadas. Infelizmente, não podemos usar a equação (7.11) para verificar se a diferença estimada entre as mulheres solteiras e as casadas é estatisticamente significante. O conhecimento dos erros-padrão de *mcasadas* e *msolteiras* não é suficiente para realizar o teste (veja Seção 4.4). O mais fácil a fazer é selecionar um desses grupos para ser o grupo base e reestimar a equação. Nada de substancial mudará, mas obteremos a estimativa necessária e seu erro-padrão de forma direta. Quando usamos as mulheres casadas como o grupo base, obtemos

$$\begin{aligned}\log(\widehat{salário}) = {} & 0{,}123 + 0{,}411\,hcasados + 0{,}198\,hsolteiros + 0{,}088\,msolteiras + \ldots, \\ & \quad (0{,}106)\ (0{,}056) \qquad\quad (0{,}058) \qquad\quad (0{,}052)\end{aligned}$$

onde, é claro, nenhum dos coeficientes ou erros-padrão não descritos sofreram alterações. A estimativa do coeficiente de *msolteiras* é, como esperado, 0,088. Agora temos um erro-padrão para acompanhar essa esti-

EXEMPLO 7.6 (continuação)

mativa. A estatística t para a hipótese nula de que não existe diferença na população entre mulheres casadas e solteiras é $t_{msolteiras} = 0{,}88/0{,}052 \approx 1{,}69$. Essa é uma evidência marginal contra a hipótese nula. Também vemos que a diferença estimada entre os homens casados e as mulheres casadas é, estatisticamente, muito significante ($t_{hcasados} = 7{,}34$).

O exemplo anterior ilustra um princípio geral para a inclusão de variáveis *dummy* que indicam grupos diferentes: se o modelo de regressão deve ter diferentes interceptos para, digamos, g grupos ou categorias, precisamos incluir $g - 1$ variáveis *dummy* no modelo, juntamente com um intercepto. O intercepto do grupo base é o intercepto global no modelo, e o coeficiente da variável *dummy* de um determinado grupo representa a diferença estimada nos interceptos entre aquele grupo e o grupo base. A inclusão de g variáveis *dummy* juntamente com um intercepto resultará na armadilha da variável *dummy*. Uma alternativa é incluir g variáveis *dummy* e excluir um intercepto global. Isso não é recomendável, pois o teste de diferenças relativas a um grupo base se tornará difícil, e alguns programas de regressão alteram a maneira como o R-quadrado é computado quando a regressão não contém um intercepto.

Nos dados sobre salários dos jogadores de beisebol encontrados no arquivo MLB1.RAW, os jogadores ocupam uma de seis posições: *pribase*, *segbase*, *terbase*, *interbase*, *jardext* ou *receptor*. Para possibilitar diferenças salariais entre as posições, com determinados defensores (*jardext*) como o grupo base, quais variáveis *dummy* você incluiria como variáveis independentes?

Incorporação de Informações Ordinais com o Uso de Variáveis *Dummy*

Suponha que gostaríamos de estimar o efeito do risco de crédito das cidades sobre as taxas de juros dos títulos públicos municipais (*TTM*). Várias instituições financeiras, como a Moody's Investment Service e a Standard and Poor's, classificam a qualidade da dívida de governos locais, na qual a classificação depende de fatores como a probabilidade de inadimplência (governos locais preferem taxas menores de juros para reduzir seus custos de empréstimos). À guisa de simplicidade, suponha que a classificação varie de zero a quatro, na qual zero é o pior risco de crédito e quatro, o melhor. Este é um exemplo de uma **variável ordinal**. Vamos chamar essa variável de *CR* por definição. A questão com a qual devemos tratar é: Como incorporamos a variável *CR* em um modelo para explicar a variável *TTM*?

Uma possibilidade é apenas incluir *CR* como incluiríamos qualquer outra variável explicativa:

$$TTM = \beta_0 + \beta_1 CR + outros\ fatores,$$

onde deliberadamente não mostramos quais são os outros fatores. Neste caso, β_1 é a mudança em pontos percentuais em *TTM* quando *CR* aumenta uma unidade, mantendo fixos todos os outros fatores. Infelizmente, é bastante difícil interpretar um aumento de uma unidade em *CR*. Sabemos o significado quantitativo de mais um ano de educação, ou de um dólar a mais gasto por aluno, mas fatores como risco de crédito, em geral, têm apenas significado ordinal. Sabemos que um *CR* de quatro é melhor que

um *CR* de três, mas será que a diferença entre quatro e três é a mesma que a diferença entre um e zero? Se não, não fará sentido assumir que um aumento de uma unidade em *CR* terá um efeito constante sobre *TTM*.

Uma abordagem melhor que podemos implementar, pois *CR* assume relativamente poucos valores, é definir variáveis *dummy* para cada valor de *CR*. Assim, definimos $CR_1 = 1$ se $CR = 1$, e, caso contrário, $CR_1 = 0$, $CR_2 = 1$ se $CR = 2$ e, caso contrário, $CR_2 = 0$, e assim por diante. Na realidade, levamos em conta o risco de crédito e o transformamos em cinco categorias. Desta forma podemos estimar o modelo

$$TTM = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + outros\ fatores. \qquad \textbf{(7.12)}$$

Seguindo nossa regra sobre inclusão de variáveis *dummy* em um modelo, incluímos quatro variáveis *dummy*, já que temos cinco categorias. A categoria aqui omitida é risco de crédito zero, e portanto ela é o grupo base. (Esta é a razão pela qual não precisamos definir uma variável *dummy* para esta categoria.) Os coeficientes são de fácil interpretação: δ_1 é a diferença em *TTM* (outros fatores fixos) entre uma cidade com um risco de crédito um e uma cidade com um risco de crédito zero: δ_2 é a diferença em *TTM* entre uma cidade com um risco de crédito dois e uma cidade com um risco de crédito zero; e assim por diante. O movimento entre os índices de risco de crédito tem efeitos diferentes, de modo que o uso de (7.12) é muito mais flexível do que simplesmente considerarmos *CR* uma variável única. Uma vez definidas as variáveis *dummy*, estimar (7.12) é simples.

No modelo (7.12), como você testaria a hipótese nula de que o risco de crédito não tem efeito sobre *TTM*?

A equação (7.12) contém o modelo com um efeito parcial constante, sendo este um caso especial. Uma maneira de escrever as três restrições que implicam um efeito parcial constante é $\delta_2 = 2\delta_1$, $\delta_3 = 3\delta_1$, e $\delta_4 = 4\delta_1$. Quando incorporamos essas restrições à equação (7.12) e a reorganizamos, obtemos $TTM = \beta_0 + \delta_1 (CR_1 + 2CR_2 + 3CR_3 + 4CR_4) + outros\ fatores$. Agora, o termo que multiplica δ_1 é simplesmente a variável original do risco de crédito, *CR*. Ao obtermos a estatística *F* para testar as restrições do efeito parcial constante, obtemos o *R*-quadrado irrestrito de (7.12) e o *R*-quadrado restrito a partir da regressão de *TTM* sobre *CR* e os outros fatores que tenhamos controlado. A estatística *F* é obtida como na equação (4.41) com $q = 3$.

EXEMPLO 7.7

(Efeitos da Atratividade Física sobre os Salários)

Hamermesh e Biddle (1994) usaram indicadores de boa aparência física em uma equação de salários. Cada pessoa da amostra foi classificada por um entrevistador quanto à aparência, utilizando cinco categorias (feia, comum, média, bonita e muito bonita). Como pouca gente se classifica nos dois extremos, os autores colocaram as pessoas em um dos três grupos para a análise de regressão, média, abaixo da média e acima da média, na qual o grupo base era a média. Utilizando os dados da Quality of Employment Survey (Pesquisa de Qualidade do Emprego) de 1977, após terem sido controladas as características de produtividade habituais, Hamermesh e Biddle estimaram uma equação para homens:

EXEMPLO 7.7 (continuação)

$$\widehat{\log(salário)} = \hat{\beta}_0 - 0{,}164\ abaixomed + 0{,}016\ acimamed + outros\ fatores$$
$$(0{,}046) \qquad (0{,}033)$$
$$n = 700,\ \bar{R}^2 = 0{,}403$$

e uma equação para mulheres:

$$\widehat{\log(salário)} = \hat{\beta}_0 - 0{,}124\ abaixomed + 0{,}035\ acimamed + outros\ fatores$$
$$(0{,}066) \qquad (0{,}049)$$
$$n = 409,\ \bar{R}^2 = 0{,}330.$$

Os outros fatores controlados na regressão incluem educação, experiência, permanência, estado civil e raça; veja Tabela 3 no artigo de Hamermesh e Biddle para uma lista mais completa. Para economizar espaço, os coeficientes das outras variáveis e do intercepto não são descritos no artigo.

Entre os homens, o artigo prevê que aqueles com aparência abaixo da média ganham cerca de 16,4% menos que os com aparência média com os mesmos fatores (inclusive educação, experiência, permanência, estado civil e raça). O efeito é estatisticamente diferente de zero, com $t = -3{,}57$. De modo semelhante, homens com aparência acima da média ganham 1,6% mais, embora o efeito não seja estatisticamente significante ($t < 0{,}5$).

Uma mulher com aparência abaixo da média ganha cerca de 12,4% menos que outra com aparência média, com $t = -1{,}88$. Como aconteceu com os homens, a estimativa do coeficiente de *acimamed* não é estatisticamente diferente de zero.

Em alguns casos, a variável ordinal assume um número muito grande de valores, de maneira que não é possível incluir uma variável *dummy* para cada valor. Por exemplo, o arquivo LAWSCH85.RAW contém dados sobre a mediana dos salários iniciais dos formados em faculdades de direito. Uma das principais variáveis explicativas é a classificação da faculdade de direito. Como cada faculdade tem uma posição diferente, evidentemente não podemos incluir uma variável *dummy* para cada posição. Se não quisermos colocar a classificação diretamente na equação, podemos dividi-la em categorias. O exemplo seguinte mostra como isso é feito.

EXEMPLO 7.8

(Efeitos da Classificação das Faculdades de Direito sobre Salários Iniciais)

Defina as variáveis *dummy*, *r11_25*, *r26_40*, *r41_60* e *r61_100*, assumindo valor unitário quando a variável *rank* cair na faixa apropriada. Definimos o grupo base como sendo os das faculdades classificadas abaixo de 100. A equação estimada é

EXEMPLO 7.8 (continuação)

$$\begin{aligned}\log(\widehat{sálário}) &= 9{,}17 + 0{,}700\,top10 + 0{,}594\,r11\_25 + 0{,}375\,r26\_40 \\ &\quad (0{,}41)\ (0{,}053)\qquad (0{,}039)\qquad (0{,}034) \\ &\quad + 0{,}263\,r41\_60 + 0{,}132\,r61\_100 + 0{,}0057\,lsat \\ &\quad (0{,}028)\qquad (0{,}021)\qquad (0{,}0031) \\ &\quad + 0{,}014\,nmdgrad + 0{,}036\log(volbib) + 0{,}0008\log(custo) \\ &\quad (0{,}074)\qquad (0{,}026)\qquad (0{,}0251) \\ n &= 136,\ R^2 = 0{,}911,\ \overline{R}^2 = 0{,}905.\end{aligned} \tag{7.13}$$

Vemos imediatamente que todas as variáveis *dummy* que definem as diferentes classificações são estatisticamente bastante significantes. A estimativa do coeficiente de *r61_100* significa que, mantendo fixos *lsat*, *nmdgrad*, *volbib* e *custo*, o salário mediano de ex-alunos formados em uma escola classificada entre as posições 61 e 100 é cerca de 13,2% mais alto do que aqueles de uma escola classificada abaixo de 100. A diferença entre as dez primeiras e as abaixo de 100 é bastante elevada. Fazendo os cálculos exatos dados na equação (7.10) temos exp(0,700) − 1 ≈ 1,014, e assim o salário mediano previsto é mais de 100% mais alto nas dez primeiras escolas do que nas abaixo de 100.

Para indicar se a divisão da classificação em diferentes grupos é um aperfeiçoamento, podemos comparar o R-quadrado ajustado de (7.13) com o R-quadrado ajustado com *rank* incluído como uma variável única: o primeiro é 0,905 e o segundo é 0,836, de modo que a flexibilidade adicional de (7.13) está garantida.

Curiosamente, quando a classificação é colocada nas categorias dadas (admitidamente arbitrárias), todas as outras variáveis tornam-se não significantes. Aliás, um teste para verificar a significância conjunta de *lsat*, *nmdgrad*, log(*volbib*) e log(*custo*) produz um p-valor de 0,055, que está no limite da significância. Quando *rank* é incluída em sua forma original, o p-valor da significância conjunta é zero até quatro casas decimais.

Um comentário final sobre este exemplo: na derivação das propriedades dos mínimos quadrados ordinários, assumimos que tínhamos uma amostra aleatória. A aplicação atual infringe essa hipótese devido à maneira como *rank* foi definida: a classificação de uma faculdade necessariamente depende da classificação das outras escolas da amostra, e portanto os dados não podem representar extrações independentes da população de faculdades de direito. Isso não causa qualquer problema mais grave, desde que o termo erro seja não-correlacionado com as variáveis explicativas.

7.4 INTERAÇÕES ENVOLVENDO VARIÁVEIS *DUMMY*

Interações entre Variáveis *Dummy*

Assim como as variáveis com significados quantitativos podem interagir em modelos de regressão, as variáveis *dummy* também podem. Vimos uma ilustração disso no Exemplo 7.6, no qual definimos quatro categorias com base em estado civil e gênero. Aliás, podemos reformular aquele modelo

adicionando um **termo de interação** entre *feminino* e *casado*, onde essas variáveis apareçam separadamente. Isso possibilita que o prêmio por ser casado dependa do gênero, como era o caso em (7.11). Com o propósito de comparação, o modelo estimado com o termo de interação *feminino-casado* é

$$\begin{aligned}\widehat{\log(salário)} &= 0{,}321 - 0{,}110\,feminino + 0{,}213\,casado \\ &\qquad (0{,}100)\quad (0{,}056)\qquad\quad (0{,}055) \\ &\quad - 0{,}301\,feminino \cdot casado + ..., \\ &\qquad (0{,}072)\end{aligned} \tag{7.14}$$

onde o restante da regressão será necessariamente idêntico a (7.11). A equação (7.14) mostra explicitamente que existe uma interação estatisticamente significante entre gênero e estado civil. Este modelo também permite obter o diferencial estimado de salários entre todos os quatro grupos, mas aqui devemos ter o cuidado de inserir a correta combinação de valores zero e um.

A definição *feminino* = 0 e *casado* = 0 corresponde ao grupo de homens solteiros, que é o grupo base, já que isso elimina *feminino*, *casado*, e *feminino-casado*. Podemos encontrar o intercepto de homens casados definindo *feminino* = 0 e *casado* = 1 em (7.14); isso produz um intercepto de 0,321 + 0,213 = 0,534, e assim por diante.

A equação (7.14) é apenas uma maneira diferente de encontrar diferenciais de salários entre todas as combinações de gênero e estado civil. Ela nos possibilita facilmente testar a hipótese nula de que o diferencial de gênero não depende do estado civil (e, de forma equivalente, que o diferencial de estado civil não depende do gênero). A equação (7.11) é mais conveniente para testarmos os diferenciais salariais entre qualquer grupo e o grupo base de homens solteiros.

EXEMPLO 7.9

(Efeitos da Utilização de Computadores nos Salários)

Krueger (1993) estima os efeitos da utilização de computadores sobre os salários. Ele define uma variável *dummy*, que chamaremos *comptrab*, igual a um se a pessoa usa um computador no trabalho. Outra variável *dummy*, *compcasa*, é igual a um se a pessoa usa um computador em casa. Utilizando 13.379 pessoas da Current Population Survey (Censo Populacional Corrente) de 1989, Krueger (1993, Tabela 4) obtém

$$\begin{aligned}\widehat{\log(salário)} &= \hat{\beta}_0 + 0{,}177\,comptrab + 0{,}070\,compcasa \\ &\qquad\qquad (0{,}009)\qquad\quad (0{,}019) \\ &\quad + 0{,}017\,comptrab \cdot compcasa + outros\,fatores. \\ &\qquad (0{,}023)\end{aligned} \tag{7.15}$$

(Os outros fatores são os padrões para regressões de salários, inclusive educação, experiência, gênero e estado civil; veja o ensaio de Krueger para a lista exata.) Krueger não descreve o intercepto porque ele não é importante; tudo que precisamos saber é que o grupo base consiste de pessoas que não usam computador em casa ou no trabalho. Vale a pena observar que o retorno estimado do uso de computador no trabalho

EXEMPLO 7.9 (continuação)

(mas não em casa) está em torno de 17,7%. (A estimativa mais precisa é de 19,4%.) De forma semelhante, as pessoas que usam o computador em casa, não no trabalho, têm um prêmio salarial em torno de 7% sobre as que não usam computador, em casa ou no trabalho. O diferencial entre os que usam o computador em ambos os locais e os que simplesmente não usam computador é de cerca de 26,4% (obtida com a adição dos três coeficientes e multiplicando o resultado por 100), ou a estimativa mais precisa de 30,2% obtida da equação (7.10).

O termo de interação em (7.15) não é estatisticamente significante, e tampouco muito grande economicamente. Entretanto, sua inclusão prejudica muito pouco a equação.

Consideração de Inclinações Diferentes

Vimos vários exemplos de como permitir diferentes interceptos para qualquer número de grupos em um modelo de regressão múltipla. Também existem casos de interação de variáveis *dummy* com variáveis explicativas que não são *dummy* para permitir uma **diferença nas inclinações**. Continuando com o exemplo salarial, suponha que queiramos verificar se o retorno da educação é o mesmo para homens e mulheres, considerando um diferencial de salários constante entre homens e mulheres (um diferencial do qual já encontramos comprovação). Para simplificar, incluímos somente educação e gênero no modelo. Que tipo de modelo leva em conta retornos diferentes em educação? Considere o modelo

$$\log(salárioh) = (\beta_0 + \delta_0 feminino) + (\beta_1 + \delta_1 feminino)educ + u. \qquad \textbf{(7.16)}$$

Se fizermos $feminino = 0$ em (7.16), veremos que o intercepto de homens é β_0, enquanto a inclinação na educação dos homens é β_1. Para as mulheres, usamos $feminino = 1$; assim, o intercepto para as mulheres será $\beta_0 + \delta_0$ e a inclinação será $\beta_1 + \delta_1$. Portanto, δ_0 mede a diferença nos interceptos entre mulheres e homens, enquanto δ_1 mede a diferença no retorno da educação entre mulheres e homens. Dois dos quatro casos dos sinais de δ_0 e δ_1 são apresentados na Figura 7.2.

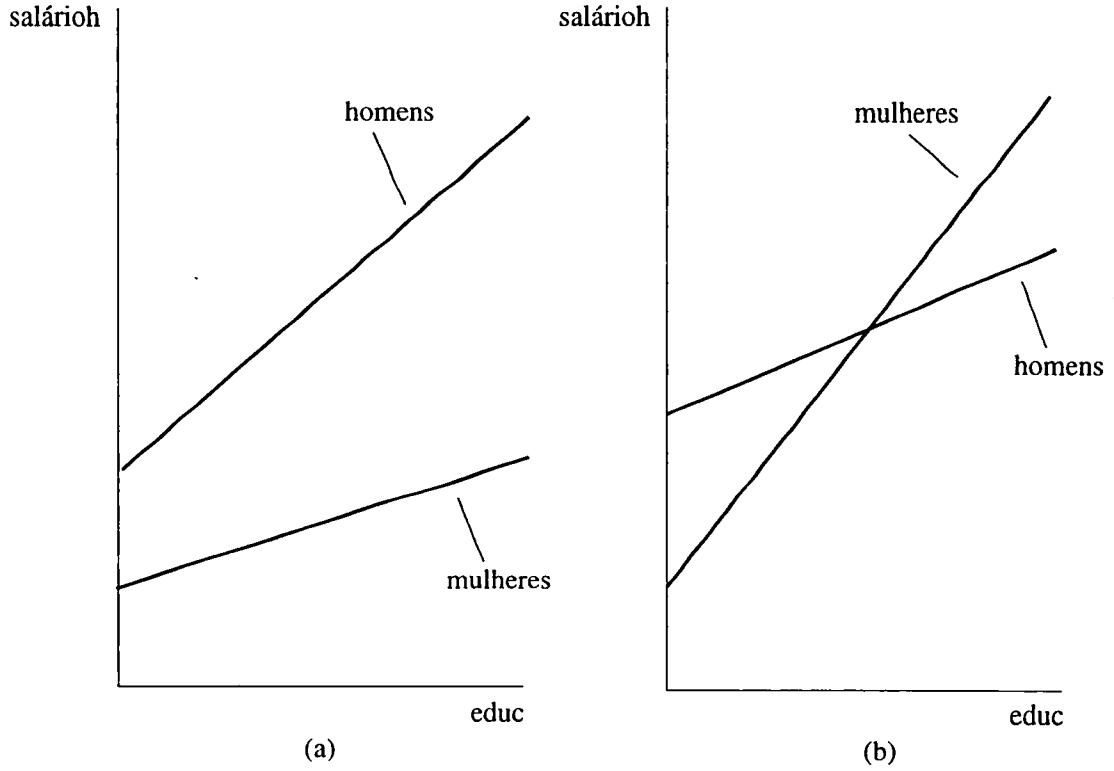
O gráfico (a) mostra o caso em que o intercepto das mulheres está abaixo do intercepto dos homens, enquanto a inclinação da linha é menor para as mulheres do que para os homens. Isso significa que as mulheres ganham menos que os homens em todos os níveis de educação, e a diferença aumenta conforme *educ* se torna maior. No gráfico (b), o intercepto das mulheres está abaixo do intercepto dos homens, mas a inclinação da educação é maior para as mulheres. Isso significa que as mulheres ganham menos que os homens em baixos níveis de educação, mas a diferença diminui conforme a educação aumenta. Em algum ponto, uma mulher ganhará mais que um homem, dado o mesmo nível de educação (e esse ponto é facilmente encontrado, dada a equação estimada).

Como podemos estimar o modelo (7.16)? Para aplicar o MQO, devemos escrever o modelo com uma interação entre *feminino* e *educ*:

$$\log(salárioh) = \beta_0 + \delta_0 feminino + \beta_1 educ + \delta_1 feminino \cdot educ + u. \qquad \textbf{(7.17)}$$

Figura 7.2

Gráficos da equação (7.16). (a) $\delta_0 < 0, \delta_1 < 0$; (b) $\delta_0 < 0, \delta_1 > 0$.



Os parâmetros agora podem ser estimados a partir da regressão de log(*salárioh*) sobre *feminino*, *educ* e *feminino·educ*. A obtenção do termo de interação é fácil com o uso de qualquer programa de regressão. Não se assuste com a natureza estranha de *feminino·educ*, que será zero para qualquer homem na amostra e igual ao nível de educação de qualquer mulher na amostra.

Uma hipótese importante é que o retorno da educação é o mesmo para mulheres e homens. Em termos do modelo (7.17), isso é declarado como H_0: $\delta_1 = 0$, o que significa que a inclinação de log(*salárioh*) em relação a *educ* é a mesma para homens e mulheres. Observe que esta hipótese não faz nenhuma restrição sobre a diferença nos interceptos, δ_0. Um diferencial de salários entre homens e mulheres é admitido nessa hipótese nula, mas ele deve ser o mesmo em todos os níveis de educação. Essa circunstância é descrita pela Figura 7.1.

Também estamos interessados na hipótese de que os salários médios são idênticos para homens e mulheres que tenham os mesmos níveis de educação. Isso significa que δ_0 e δ_1 devem *ambos* ser zero sob a hipótese nula. Na equação (7.17) precisamos usar um teste F para testar H_0: $\delta_0 = 0, \delta_1 = 0$. No modelo com apenas uma diferença de interceptos, rejeitamos essa hipótese, pois H_0: $\delta_0 = 0$ é completamente rejeitada contra H_1: $\delta_0 < 0$.

EXEMPLO 7.10

(Equação do Log dos Salários-Hora)

Adicionemos termos quadráticos de experiência e permanência à equação (7.17):

$$\begin{aligned}\widehat{\log(salário h)} &= 0{,}389 - 0{,}227\,feminino + 0{,}082\,educ \\ &\quad (0{,}119)\ (0{,}168)\qquad (0{,}008) \\ &\quad - 0{,}0056\,feminino\cdot educ + 0{,}029\,exper - 0{,}00058\,exper^2 \\ &\quad (0{,}0131)\qquad\qquad (0{,}005)\qquad (0{,}00011) \\ &\quad + 0{,}032\,perm - 0{,}00059\,perm^2 \\ &\quad (0{,}007)\qquad (0{,}00024) \\ n &= 526,\ R^2 = 0{,}441.\end{aligned} \tag{7.18}$$

O retorno estimado da educação dos homens nesta equação é 0,082, ou 8,2%. Para mulheres, o retorno é 0,082 − 0,0056 = 0,0764, ou cerca de 7,6%. A diferença, −0,56%, ou pouco mais de meio ponto percentual a menos para as mulheres, não é economicamente grande nem estatisticamente significante: a estatística t é −0,0056/0,0131 ≈ −0,43. Assim, concluímos que não há comprovação contra a hipótese de que o retorno da educação seja o mesmo para homens e mulheres.

O coeficiente de *feminino*, embora permaneça economicamente grande, não é mais significante aos níveis convencionais ($t = -1{,}35$). Seu coeficiente e a estatística t na equação sem a interação eram −0,297 e −8,25, respectivamente [veja a equação (7.9)]. Devemos agora concluir que não existe evidência estatisticamente significante de salários mais baixos para mulheres nos mesmos níveis de *educ*, *exper* e *perm*? Isso seria um erro grave. Como adicionamos a interação *feminino*·*educ* à equação, o coeficiente de *feminino* é agora estimado com muito menos precisão do que na equação (7.9): o erro-padrão aumentou em quase cinco vezes (0,168/0,036 ≈ 4,67). A razão disto é que *feminino* e *feminino*·*educ* são altamente correlacionados na amostra. Neste exemplo, existe uma maneira proveitosa de pensar sobre a multicolinearidade: na equação (7.17) e na equação mais geral estimada em (7.18), δ_0 mede o diferencial de salários entre mulheres e homens quando *educ* = 0. Como não existe ninguém na amostra com anos de educação sequer próximo de zero, não é surpreendente que temos muito trabalho para estimar o diferencial em *educ* = 0 (tampouco o diferencial em zero anos de educação é muito informativo). Mais interessante seria estimar o diferencial por gênero no, digamos, nível médio de educação da amostra (cerca de 12,5). Para fazer isso temos que substituir *feminino*·educ por *feminino*·(*educ* − 12,5) e computar novamente a regressão; isso muda apenas o coeficiente de *feminino* e seu erro-padrão. (Veja o Exercício 7.15.)

Se computarmos a estatística F de H_0: $\delta_0 = 0, \delta_1 = 0$, obteremos $F = 34{,}33$, que é um enorme valor para uma variável aleatória F com numerador $gl = 2$ e denominador $gl = 518$: o p-valor é zero até quatro casas decimais. No final, preferimos o modelo (7.9), que considera um diferencial de salários constante entre mulheres e homens.

Como você ampliaria o modelo estimado em (7.18) para possibilitar que o retorno de *perm* diferencie por gênero?

Como um exemplo mais complicado envolvendo interações, examinamos agora os efeitos da raça e da composição racial das cidades sobre os salários dos jogadores de beisebol da liga principal desse esporte nos Estados Unidos.

EXEMPLO 7.11

(Efeitos da Raça sobre os Salários dos Jogadores de Beisebol)

A equação seguinte é estimada para os 330 jogadores de beisebol da liga principal das cidades onde estão disponíveis estatísticas sobre a composição racial. As variáveis *negro* e *hispan* são indicadores binários dos jogadores individuais. (O grupo base é formado pelos jogadores brancos.) A variável *porcnegro* é a percentagem de negros na cidade da equipe, enquanto *porchisp* é a percentagem de hispânicos. As outras variáveis indicam aspectos da produtividade e da longevidade dos jogadores. Neste caso, estamos interessados no efeito da raça após termos controlado esses fatores.

Além de termos incluído *negro* e *hispan* na equação, adicionamos as interações *negro·porcnegro* e *hispan·porchisp*. A equação estimada é

$$\begin{aligned}\widehat{\log(salário)} &= \underset{(2,18)}{10,34} + \underset{(0,0129)}{0,0673}\, anos + \underset{(0,0034)}{0,0089}\, jogosano \\ &+ \underset{(0,00151)}{0,00095}\, rebmed + \underset{(0,0164)}{0,0146}\, hrunano + \underset{(0,0076)}{0,0045}\, rbisyr \\ &+ \underset{(0,0046)}{0,0072}\, runsano + \underset{(0,0021)}{0,0011}\, perccap + \underset{(0,0029)}{0,0075}\, porcest \\ &- \underset{(0,125)}{0,198}\, negro - \underset{(0,153)}{0,190}\, hispan + \underset{(0,0050)}{0,0125}\, negro\cdot porcnegro \\ &+ \underset{(0,0098)}{0,0201}\, hispan\cdot porchisp, \quad n = 330,\ R^2 = 0,638.\end{aligned} \tag{7.19}$$

Primeiro, devemos verificar se as quatro variáveis raciais, *negro*, *hispan*, *negro·porcnegro* e *hispan·porchisp* são conjuntamente significantes. Usando os mesmos 330 jogadores, o R-quadrado quando as quatro variáveis raciais são eliminadas é 0,626. Como existem quatro restrições e $gl = 330 - 13$ no modelo sem restrições, a estatística F está em torno de 2,63, o que produz um p-valor de 0,034. Assim, essas variáveis são conjuntamente significantes ao nível de 5% (embora não o sejam ao nível de 1%).

Como interpretaremos os coeficientes das variáveis raciais? Na discussão seguinte, todos os fatores de produtividade são mantidos fixos. Primeiro, considere o que acontece com jogadores negros, mantendo fixo *porchisp*. O coeficiente $-0,198$ de *negro* literalmente significa que se um jogador negro estiver em uma cidade onde não haja negros (*porcnegro* = 0), então o jogador negro ganhará cerca de 19,8% menos do que um jogador branco nas mesmas condições. Na medida em que *porcnegro* aumenta — o que significa que a população branca diminui, já que *porchisp* é mantida fixo — os salários dos negros aumentam comparado aos dos brancos. Em uma cidade com 10% de negros, log(*salário*) dos negros comparado com o dos brancos é $-0,198 + 0,0125(10) = -0,073$, ou seja, os salários dos negros serão cerca de 7,3% menores

EXEMPLO 7.11 (continuação)

que os dos brancos em tal cidade. Quando *porcnegro* = 20, os negros ganham cerca de 5,2% mais que os brancos. A maior percentagem de negros em uma cidade está em torno de 74% (Detroit).

De forma semelhante, os hispânicos ganham menos que os brancos em cidades com um baixo percentual de hispânicos. Mas podemos facilmente encontrar o valor de *porchisp* que torna o diferencial entre brancos e hispânicos igual a zero: ele deve ser $-0,190 + 0,0201\ porchisp = 0$, o que produz $porchisp \approx 9,45$. Em cidades nas quais a percentagem de hispânicos for menor que 9,45%, é possível prever que os hispânicos ganharão menos que os brancos (para uma determinada população de negros), e o oposto é verdadeiro se o número de hispânicos estiver acima de 9,45%. Doze das vinte e duas cidades representadas na amostra possuem população hispânica menor que 6% da população total. A maior percentagem de hispânicos está em torno de 31%.

Como interpretar esses resultados? Não podemos simplesmente alegar que existe discriminação contra negros e hispânicos, porque as estimativas indicam que os brancos ganham menos que os negros e os hispânicos em cidades densamente povoadas por minorias. A importância da composição racial de uma cidade sobre os salários pode ser devida às preferências dos jogadores: talvez os melhores jogadores negros vivam em cidades com mais negros e os melhores jogadores hispânicos tendam a viver em cidades com maior concentração de hispânicos. As estimativas em (7.19) nos possibilitam determinar a presença de alguma relação, mas não temos condições de fazer a distinção entre essas duas hipóteses.

Verificação de Diferenças nas Funções de Regressão entre Grupos

Os exemplos anteriores ilustram que a interação de variáveis *dummy* com outras variáveis independentes pode ser uma ferramenta poderosa. Algumas vezes, queremos testar a hipótese nula de que duas populações, ou grupos, seguem a mesma função de regressão, contra a hipótese alternativa de que uma ou mais das inclinações diferem entre os grupos. Também veremos exemplos disso no Capítulo 13, quando discutiremos o agrupamento de diferentes cortes transversais ao longo do tempo.

Suponha que queiramos testar se o mesmo modelo de regressão descreve a nota média no curso superior de atletas universitários masculinos e femininos. A equação é

$$nmgradac = \beta_0 + \beta_1 sat + \beta_2 emperc + \beta_3 tothrs + u,$$

onde *sat* é a nota obtida no exame de ingresso em curso superior, *emperc* é o percentil da classificação no ensino médio, e *tothrs* é o total de horas do curso superior. Sabemos que para considerar uma diferença nos interceptos podemos incluir uma variável *dummy* para masculino ou feminino. Se quisermos que qualquer uma das inclinações dependa do gênero, simplesmente fazemos a interação da variável apropriada com, digamos, *feminino*, e a incluímos na equação.

Se estivermos interessados em verificar se existe *qualquer* diferença entre homens e mulheres, então devemos admitir um modelo no qual o intercepto e todas as inclinações possam ser diferentes entre os grupos:

$$\begin{aligned} nmgradac = {} & \beta_0 + \delta_0 feminino + \beta_1 sat + \delta_1 feminino \cdot sat + \beta_2 emperc \\ & + \delta_2 feminino \cdot emperc + \beta_3 tothrs + \delta_3 feminino \cdot tothrs + u. \end{aligned} \qquad \textbf{(7.20)}$$

O parâmetro δ_0 é a diferença nos interceptos entre mulheres e homens, δ_1 é a diferença de inclinações em relação a *sat* entre mulheres e homens, e assim por diante. A hipótese nula de que *nmgradac* segue o mesmo modelo para homens e mulheres é escrita como

$$H_0: \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0. \tag{7.21}$$

Se um dos δ_j for diferente de zero, então os modelos são diferentes para homens e mulheres.

Utilizando os dados do arquivo GPA3.RAW, o modelo completo é estimado como

$$\begin{aligned} \widehat{nmgradac} = {} & 1{,}48 - 0{,}353\,feminino + 0{,}0011\,sat + 0{,}00075\,feminino \cdot sat \\ & (0{,}21)\ (0{,}411) \qquad (0{,}0002) \qquad (0{,}00039) \\ & - 0{,}0085\,emperc - 0{,}00055\,feminino \cdot emperc + 0{,}0023\,tothrs \\ & (0{,}0014) \qquad (0{,}00316) \qquad (0{,}0009) \\ & - 0{,}00012\,feminino \cdot tothrs \\ & (0{,}00163) \\ & n = 366,\ R^2 = 0{,}406,\ \overline{R}^2 = 0{,}394. \end{aligned} \tag{7.22}$$

A variável *dummy feminino* e todos os termos de interação não são muito significantes; somente a interação *feminino·sat* tem uma estatística *t* próxima de dois. Entretanto, sabemos que não devemos confiar nas estatísticas *t* individuais para testar uma hipótese conjunta como (7.21). Para computar a estatística *F* devemos estimar o modelo restrito, que resulta da eliminação de *feminino* e de todas as interações; isso produz um R^2 (o R^2 restrito) em torno de 0,352, de modo que a estatística *F* está em torno de 8,14; o *p*-valor é zero até cinco casas decimais, o que nos leva a rejeitar completamente (7.21). Assim, os modelos que especificam *nmgradac* de atletas masculinos e femininos são diferentes, embora cada termo em (7.22), que permitem que homens e mulheres sejam diferentes, sejam individualmente não significantes ao nível de 5%.

O grande erro-padrão da variável *feminino* e os termos de interação tornam difícil dizer com precisão como diferem homens e mulheres. Precisamos ter muito cuidado na interpretação da equação (7.22), pois, na obtenção das diferenças entre homens e mulheres, os termos de interação devem ser levados em conta. Se olharmos somente a variável *feminino*, concluiremos erroneamente que *nmgradac* é cerca de 0,353 menor para mulheres do que para homens, mantendo fixos os outros fatores. Esta é a diferença estimada somente quando *sat*, *emperc* e *tothrs* são definidas como zero, o que não é um cenário interessante. Com *sat* = 1.100, *emperc* = 10 e *tothrs* = 50, a *diferença* prevista entre uma mulher e um homem é $-0{,}353 + 0{,}00075\,(1.100) - 0{,}00055(10) - 0{,}00012(50) \approx 0{,}461$. Ou seja, é possível prever que a atleta feminina tem *nmgradac* quase meio ponto mais alta que um atleta masculino nas mesmas condições.

Em um modelo com três variáveis, *sat*, *emperc* e *tothrs*, é muito simples adicionar todas as interações para testar diferenças entre grupos. Em alguns casos, muito mais variáveis explicativas estão envolvidas, e portanto é conveniente termos uma maneira diferente de computar a estatística. A soma dos resíduos quadrados da estatística *F* pode ser computada facilmente mesmo quando muitas variáveis independentes estão envolvidas.

No modelo geral com k variáveis explicativas e um intercepto, suponha que temos dois grupos, que chamaremos de $g = 1$ e $g = 2$. Gostaríamos de verificar se o intercepto e todas as inclinações são os mesmos nos dois grupos. Escreva o modelo como

$$y = \beta_{g,0} + \beta_{g,1}x_1 + \beta_{g,2}x_2 + ... + \beta_{g,k}x_k + u, \qquad \textbf{(7.23)}$$

para $g = 1$ e $g = 2$. A hipótese de que cada beta em (7.23) é o mesmo nos dois grupos envolve $k + 1$ restrições (no exemplo de *nmgradac*, $k + 1 = 4$). O modelo sem restrições, que pode ser entendido como tendo uma variável *dummy* de grupo e k termos de interação, além do intercepto e das próprias variáveis, tem $n - 2(k + 1)$ graus de liberdade. [No exemplo da *nmgradac*, $n - 2(k + 1) = 366 - 2(4) = 358$.] Até aqui, não há nenhuma novidade. A percepção básica é que a soma dos resíduos quadrados do modelo sem restrições pode ser obtida de duas regressões *separadas*, uma para cada grupo. Seja SQR_1 a soma dos resíduos quadrados obtida ao estimar (7.23) para o primeiro grupo; isso envolve n_1 observações. Seja SQR_2 a soma dos resíduos quadrados obtida ao estimar o modelo usando o segundo grupo (n_2 observações). No exemplo anterior, se o grupo 1 for de mulheres, $n_1 = 90$ e $n_2 = 276$. Agora, a soma dos resíduos quadrados do modelo sem restrições é simplesmente $SQR_{ir} = SQR_1 + SQR_2$. A soma dos resíduos quadrados restrita é somente a SQR do agrupamento dos grupos e da estimativa de uma única equação, digamos SQR_r. Uma vez calculados esses termos, computamos a estatística F da forma habitual:

$$F = \frac{[SQR_r - (SQR_1 + SQR_2)]}{SQR_1 + SQR_2} \cdot \frac{[n - 2(k + 1)]}{k + 1} \qquad \textbf{(7.24)}$$

onde n é o número *total* de observações. Esta estatística F específica é usualmente chamada em econometria de **estatística de Chow**. Como o teste de Chow é apenas um teste F, ele só é válido sob homoscedasticidade. Em particular, sob a hipótese nula, as variâncias dos erros dos dois grupos devem ser iguais. Como sempre, a normalidade não é necessária para a análise assimptótica.

Para aplicarmos a estatística de Chow no exemplo de *nmgradac*, precisamos da SQR da regressão que reuniu os grupos: ela é $SQR_r = 85{,}515$. A SQR das 90 mulheres na amostra é $SQR_1 = 19{,}603$ e a SQR dos homens é $SQR_2 = 58{,}752$. Portanto, $SQR_{ir} = 19{,}603 + 58{,}752 = 78{,}355$. A estatística F é $[(85{,}515 - 78{,}355)/78{,}355](358/4) \approx 8{,}18$; naturalmente, sujeito ao erro de arredondamento, isso é o que obtemos usando a forma R-quadrado do teste nos modelos com e sem os termos de interação. (Uma advertência: não existe uma forma R-quadrado simples do teste se regressões separadas forem estimadas para cada grupo; a forma R-quadrado do teste poderá ser usada somente se tiverem sido incluídas interações para criar o modelo sem restrições.)

Uma limitação importante do teste de Chow, independentemente do método usado para implementá-lo, é a hipótese nula não permitir nenhuma diferença entre os grupos. Em muitos casos, é mais interessante considerar uma diferença nos interceptos entre os grupos e depois verificar as diferenças das inclinações; vimos uma ilustração disso na equação salarial no Exemplo 7.10. Há duas maneiras de fazermos com que os interceptos difiram sob a hipótese nula. Uma delas é incluir a *dummy* do grupo e todos os termos de interação, como na equação (7.22), mas apenas testar a significância conjunta dos termos de interação. A segunda é calcular uma estatística F como na equação (7.24), mas onde a soma dos quadrados restrita, chamada "SQR_r" na equação (7.24), é obtida pela regressão que permite somente um deslocamento do intercepto. Em outras palavras, computamos uma regressão agrupada e apenas incluímos as variáveis *dummy* que distinguem os dois grupos. No exemplo da nota média do

curso superior, fazemos a regressão de *nmgradac* sobre *feminino*, *sat*, *emperc* e *tothrs*, usando os dados dos alunos-atletas femininos e masculinos. No exemplo de *nmgradac*, usamos o primeiro método, e assim a hipótese nula é H_0: $\delta_1 = 0, \delta_2 = 0, \delta_3 = 0$ na equação (7.20). (δ_0 não é restrita sob a hipótese nula.) A estatística F para essas três restrições está em torno de 1,53, o que produz um p-valor igual a 0,205. Portanto, não rejeitamos a hipótese nula.

A impossibilidade de rejeitar a hipótese de que os parâmetros que multiplicam os termos de interação são todos zero sugere que o melhor modelo permite somente uma diferença de interceptos:

$$\begin{aligned} \widehat{nmgradac} &= 1{,}39 + 0{,}310\,feminino + 0{,}0012\,sat - 0{,}0084\,emperc \\ &\quad (0{,}18)\ (0{,}059)\qquad (0{,}0002)\qquad (0{,}0012) \\ &\quad + 0{,}0025\,tothrs \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (7.25) \\ &\quad (0{,}0007) \\ n &= 366,\ R^2 = 0{,}398,\ \overline{R}^2 = 0{,}392. \end{aligned}$$

Os coeficientes das inclinações em (7.25) estão próximos daqueles do grupo base (homens) em (7.22); a eliminação das interações altera pouca coisa. Porém, *feminino* em (7.25) é altamente significativa: sua estatística t está acima de cinco, sugerindo que, em determinados níveis de *sat*, *emperc* e *tothrs*, uma atleta mulher tem uma *nmgradac* prevista que é 0,31 pontos mais alta que a de um atleta homem. Essa é, de fato, uma diferença importante.

7.5 UMA VARIÁVEL DEPENDENTE BINÁRIA: O MODELO DE PROBABILIDADE LINEAR

Até agora, aprendemos bastante sobre as propriedades e a aplicabilidade do modelo de regressão linear múltipla. Nas últimas seções, estudamos como podemos incorporar informações qualitativas, por exemplo, variáveis explicativas em um modelo de regressão múltipla, por meio do uso de variáveis independentes binárias. Em todos os modelos vistos até agora, a variável dependente y teve um significado *quantitativo* (por exemplo, y é um montante em dólares, uma pontuação em um teste, uma percentagem, ou seus logs). O que acontece se quisermos usar regressão múltipla para *explicar* um evento qualitativo?

No caso mais simples o evento que gostaríamos de explicar, e que aparece com muita freqüência na prática, é um resultado binário. Em outras palavras, nossa variável dependente, y, assume somente um dos dois valores: zero ou um. Por exemplo, y pode ser definido para indicar se um adulto concluiu o ensino médio; y pode indicar se um aluno do curso superior usou drogas ilegais durante determinado ano escolar; ou y pode indicar se uma empresa foi absorvida por outra durante determinado ano. Em cada um desses exemplos, podemos definir que $y = 1$ represente um dos resultados e $y = 0$, o outro.

Isso significaria escrever um modelo de regressão múltipla, tal como

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \qquad (7.26)$$

quando y for uma variável binária? Como y pode assumir somente dois valores, β_j não pode ser interpretado como a mudança em y devido ao aumento de uma unidade em x_j, mantendo fixos todos os

outros fatores: y somente muda de zero para um ou de um para zero. No entanto, os coeficientes β_j ainda têm interpretações úteis. Se assumirmos que a hipótese de média condicional zero RLM.3 é válida, isto é, $E(u|x_1,...,x_k) = 0$, então teremos, como sempre,

$$E(y|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k,$$

onde $\boldsymbol{x}$ é uma forma abreviada que representa todas as variáveis explicativas.

O ponto principal é que, quando y é uma variável binária assumindo os valores zero e um, é sempre verdade que $P(y = 1|\boldsymbol{x}) = E(y|\boldsymbol{x})$: a probabilidade de "sucesso" — isto é, a probabilidade de que $y = 1$ — é a mesma do valor esperado de y. Assim, temos a importante equação

$$P(y = 1|\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k, \qquad \textbf{(7.27)}$$

que mostra a probabilidade de sucesso, digamos, $p(\boldsymbol{x}) = P(y = 1|\boldsymbol{x})$, uma função linear de x_j. A equação (7.27) é um exemplo de modelo de resposta binária, e $P(y = 1|\boldsymbol{x})$ também é chamado de **probabilidade de resposta**. (Trataremos de outros modelos de resposta binária no Capítulo 17.) Como a soma das probabilidades deve ser um, $P(y = 0|\boldsymbol{x}) = 1 - P(y = 1|\boldsymbol{x})$ também é uma função linear de x_j.

O modelo de regressão linear múltipla com uma variável dependente binária é chamado de **modelo de probabilidade linear (MPL)** porque a probabilidade de resposta é linear nos parâmetros β_j. No MPL, β_j mede a mudança na probabilidade de sucesso quando x_j muda, mantendo fixos os outros fatores:

$$\Delta P(y = 1|\boldsymbol{x}) = \beta_j \Delta x_j. \qquad \textbf{(7.28)}$$

Com isso em mente, o modelo de regressão múltipla pode nos permitir estimar o efeito de diversas variáveis explicativas sobre eventos qualitativos. A mecânica do MQO é a mesma de antes.

Se escrevermos a equação estimada como

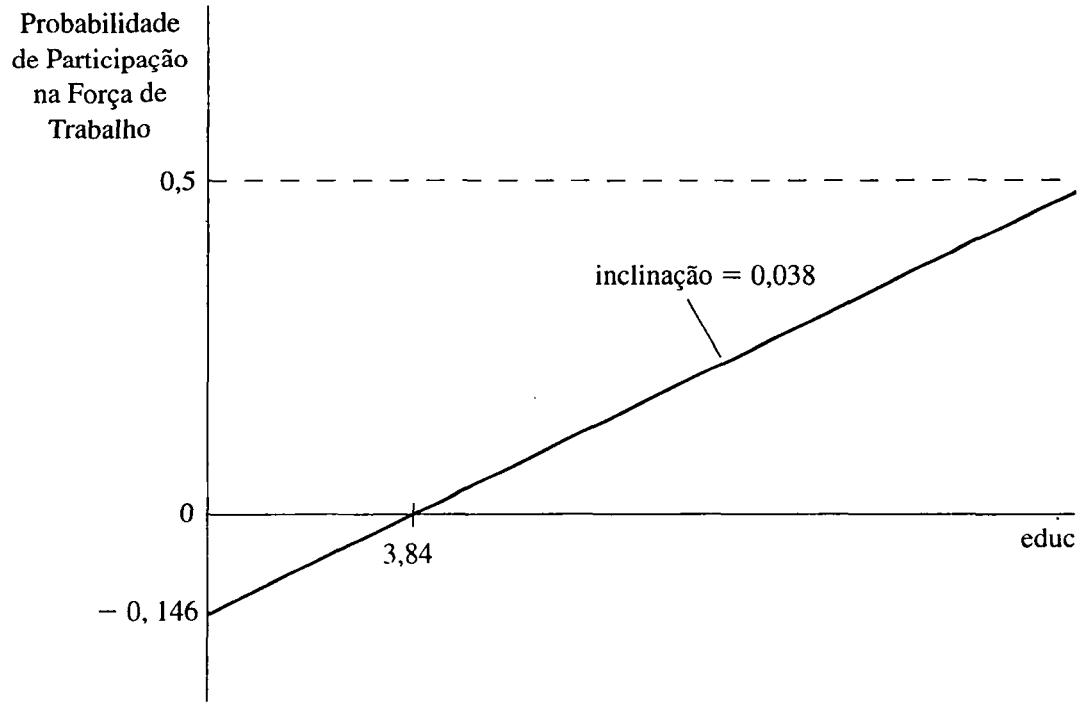
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_k x_k,$$

temos que nos lembrar que $\hat{y}$ é a probabilidade de sucesso prevista. Portanto, $\hat{\beta}_0$ é a probabilidade de sucesso prevista quando cada x_j é definido como zero, o que pode, ou não, ser interessante. O coeficiente de inclinação $\hat{\beta}_1$ mede a mudança prevista na probabilidade de sucesso quando x_1 aumenta em uma unidade.

Para interpretarmos corretamente um modelo de probabilidade linear, precisamos saber o que constitui um "sucesso". Assim, é uma boa idéia dar à variável dependente um nome que descreva o evento $y = 1$. Como exemplo, definamos *naft* ("na força de trabalho") como uma variável binária indicando a participação na força de trabalho de uma mulher casada, durante 1975: *naft* = 1 se a mulher informar ter trabalhado com remuneração fora de casa em algum período do ano, e zero, caso contrário. Assumimos que a participação na força de trabalho depende de outras fontes de renda, inclusive a renda do marido (*nesprend*, expressa em milhares de dólares), anos de estudo (*educ*), experiência anterior no mercado de trabalho (*exper*), *idade*, número de filhos menores de seis anos (*crianmed6*) e número de filhos entre 6 e 18 anos (*crianma6*). Utilizando os dados de Mroz (1987), estimamos o seguinte modelo de probabilidade linear, no qual 428 das 753 mulheres da amostra informam terem estado na força de trabalho em algum período do ano de 1975:

Figura 7.3

Relação estimada entre a probabilidade de estar na força de trabalho e anos de educação, com outras variáveis explicativas fixas.



$$\widehat{naft} = 0{,}586 - 0{,}0034\ nesprend + 0{,}038\ educ + 0{,}039\ exper$$
$$(0{,}154)\quad (0{,}0014)\qquad (0{,}007)\qquad (0{,}006)$$
$$- 0{,}00060\ exper^2 - 0{,}016\ idade - 0{,}262\ crianmed6 + 0{,}0130\ crianma6 \qquad \textbf{(7.29)}$$
$$(0{,}00018)\qquad (0{,}002)\qquad (0{,}034)\qquad (0{,}0132)$$
$$n = 753,\ R^2 = 0{,}264.$$

Usando as estatísticas *t* habituais, todas as variáveis em (7.29) exceto *crianma6* são estatisticamente significantes, e todas as variáveis significantes têm os efeitos que esperaríamos baseados na teoria econômica (ou no bom senso).

Para interpretar as estimativas, devemos nos lembrar que uma alteração na variável independente muda a probabilidade de que *naft* = 1. Por exemplo, o coeficiente de *educ* significa que, tudo o mais em (7.29) mantido fixo, mais um ano de educação, aumenta a probabilidade de participação na força de trabalho em 0,038. Se interpretarmos essa equação literalmente, mais dez anos de educação aumentarão a probabilidade de estar na força de trabalho em 0,038(10) = 0,38, o que é um aumento bastante grande em uma probabilidade. A relação entre a probabilidade de participação na força de trabalho e *educ* está traçada na Figura 7.3. As outras variáveis independentes são fixadas nos valores *nesprend* = 50, *exper* = 5, *idade* = 30, *crianmed6* = 1 e *crianma6* = 0, para fins ilustrativos. A probabilidade prevista é negativa até que o nível de educação iguale 3,84 anos. Isso não deve causar muita preocupação

porque, na amostra, nenhuma mulher tem menos de cinco anos de estudo. O nível de educação mais alto informado é de 17 anos, e isso leva a uma probabilidade prevista de 0,5. Se definíssemos as outras variáveis independentes com diferentes valores, a gama de probabilidades previstas se alteraria. Contudo, o efeito marginal de mais um ano de educação na probabilidade de participação na força de trabalho será sempre 0,038.

O coeficiente de *nesprend* sugere que, se $\Delta nesprend = 10$ (o que significa um aumento de 10.000 dólares), a probabilidade de que uma mulher esteja na força de trabalho diminui em 0,034.

Esse não é um efeito especialmente grande, considerando que um aumento na renda de 10.000 dólares é bastante significativo em termos de dólares de 1975. A experiência foi incluída como um termo quadrático para possibilitar que o efeito da experiência anterior tenha um efeito decrescente na probabilidade de participação na força de trabalho. Mantendo fixos outros fatores, a mudança estimada na probabilidade será aproximadamente $0,039 - 2(0,0006)exper = 0,039 - 0,0012\ exper$. O ponto no qual a experiência anterior não tem efeito sobre a probabilidade de participação na força de trabalho é $0,039/0,0012 = 32,5$, que é um alto nível de experiência: somente 13 das 753 mulheres na amostra têm mais de 32 anos de experiência.

Ao contrário do número de filhos mais velhos, o número de filhos mais novos tem um enorme impacto na participação na força de trabalho. Ter mais um filho com menos de seis anos de idade reduz a probabilidade de participação na força de trabalho em $-0,262$, nos níveis dados das outras variáveis. Na amostra, pouco menos de 20% das mulheres têm pelo menos um filho nessa faixa de idade.

Este exemplo ilustra o quanto é fácil estimar e interpretar os modelos de probabilidade linear, mas também destaca algumas de suas deficiências. Primeiro, é fácil verificar que, se agregarmos certas combinações de valores das variáveis independentes em (7.29), podemos obter previsões menores que zero ou maiores que um. Como estamos falando de probabilidades previstas, e probabilidades devendo estar entre zero e um, isso pode ser um pouco complicado. Por exemplo, qual seria o sentido de prever que uma mulher está na força de trabalho com uma probabilidade de $-0,10$? Aliás, das 753 mulheres na amostra, 16 dos valores estimados usando (7.29) são menores que zero, e 17 dos valores estimados são maiores que um.

Um problema relacionado é que a probabilidade não pode ser linearmente relacionada com as variáveis independentes em todos os seus possíveis valores. Por exemplo, a equação (7.29) prevê que o efeito de passar de zero filho para um filho menor de seis anos reduz a probabilidade de trabalhar em 0,262. Essa também é a redução se a mulher passar de um filho para dois. Pareceria mais realista que o primeiro filho reduzisse a probabilidade em grande escala, enquanto os filhos subseqüentes tivessem um efeito marginal menor. De fato, quando levada ao extremo, a equação (7.29) sugere que passar de zero para quatro filhos reduz a probabilidade de trabalhar em $\Delta \hat{nafft} = 0,262(\Delta crianmed6) = 0,262(4) = 1,048$, o que é impossível.

Mesmo com esses problemas, o modelo de probabilidade linear é útil e freqüentemente aplicado em economia. Normalmente, ele funciona bem com os valores das variáveis independentes que estejam próximos das médias na amostra. No exemplo da participação na força de trabalho, não existe nenhuma mulher na amostra que tenha quatro filhos menores de seis anos; aliás, somente três mulheres têm três filhos pequenos. Mais de 96% das mulheres não têm filhos ou têm apenas um, e assim provavelmente deveríamos restringir nossa atenção neste caso, quando interpretarmos a equação estimada.

Probabilidades previstas fora do intervalo da unidade são um pouco problemáticas quando queremos fazer previsões, mas muito raramente esse é o ponto central de uma análise. Normalmente, queremos saber o efeito *ceteris paribus* de certas variáveis sobre a probabilidade.

Devido à natureza binária de y, o modelo de probabilidade linear infringe uma das hipóteses de Gauss-Markov. Quando y é uma variável binária, sua variância, condicional em $\boldsymbol{x}$, é

$$\text{Var}(y|\boldsymbol{x}) = p(\boldsymbol{x})[1 - p(\boldsymbol{x})], \tag{7.30}$$

onde $p(\boldsymbol{x})$ é uma forma abreviada da probabilidade de sucesso: $p(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$.

Isso significa que, com exceção do caso em que a probabilidade não depende de qualquer das variáveis independentes, *deve* haver heteroscedasticidade no modelo de probabilidade linear. Sabemos, do Capítulo 3, que isso não causa viés nos estimadores MQO de β_j. Entretanto, sabemos também, dos Capítulos 4 e 5, que a homoscedasticidade é crucial para justificar as estatísticas t e F habituais, mesmo em amostras grandes. Como os erros-padrão em (7.29) não são, de forma geral, válidos, devemos usá-los com cuidado. Mostraremos como corrigir os erros-padrão quanto à heteroscedasticidade no Capítulo 8. Em muitas aplicações, as estatísticas MQO habituais não ficam muito distorcidas, e ainda é aceitável no trabalho aplicado apresentar uma análise MQO padrão de um modelo de probabilidade linear.

EXEMPLO 7.12

(Um Modelo de Probabilidade Linear de Prisões)

Façamos *pris86* ser uma variável binária igual à unidade se um homem foi preso durante o ano de 1986, e zero caso contrário. A população é um grupo de homens na Califórnia nascidos em 1960 ou 1961 que tenham sido presos pelo menos uma vez antes de 1986. Um modelo de probabilidade linear para descrever *pris86* é

$$pris86 = \beta_0 + \beta_1 pcond + \beta_2 sentmed + \beta_3 temptot + \beta_4 ptemp86 + \beta_5 empr86 + u,$$

onde *pcond* é a proporção de condenações anteriores, *sentmed* é a duração média da sentença por condenações anteriores (em meses), *temptot* é o número de meses passado na prisão desde os 18 anos de idade antes de 1986, *ptemp86* é o número de meses passados na prisão durante 1986, e *empr86* é o número de trimestres (0 a 4) que o homem esteve legalmente empregado em 1986.

Os dados que usamos estão no arquivo CRIME1.RAW, o mesmo conjunto de dados usado para o Exemplo 3.5. Neste caso, usamos uma variável dependente binária, porque somente 7,2% dos homens da amostra foram presos mais de uma vez. Cerca de 27,7% foram presos pelo menos uma vez durante 1986. A equação estimada é

$$\begin{aligned} \widehat{pris86} = 0{,}441 &- 0{,}162\, pcond + 0{,}0061\, sentmed - 0{,}0023\, temptot \\ &(0{,}017)\quad (0{,}021)\qquad (0{,}0065)\qquad (0{,}0050) \\ &- 0{,}022\, ptemp86 - 0{,}043\, empr86 \\ &(0{,}005)\qquad (0{,}005) \\ n = 2.725,\ &R^2 = 0{,}0474. \end{aligned} \tag{7.31}$$

O intercepto 0,441 é a probabilidade esperada de prisão de alguém que não tenha sido condenado (e portanto *pcond* e *sentmed* são ambas zero), não tenha passado tempo na prisão desde a idade de 18 anos, não tenha passado tempo na prisão em 1986, e esteve desempregado durante o ano inteiro. As variáveis *sentmed* e *temptot* não são significantes tanto individual como conjuntamente (o teste F produz p-valor = 0,347), e *sentmed* terá um sinal antiintuitivo se sentenças mais longas supostamente desencorajarem a criminalidade. Grogger (1991), utilizando um conjunto ampliado desses dados e métodos econométricos

EXEMPLO 7.12 (continuação)

diferentes, verificou que *temptot* tem um efeito positivo estatisticamente significante sobre as prisões e concluiu que *temptot* é um indicador do capital humano construído em torno de atividades criminais.

O aumento da probabilidade de condenação não reduz a probabilidade de prisões, mas precisamos ser cuidadosos quando interpretarmos a magnitude do coeficiente. A variável *pcond* é uma proporção entre zero e um; assim, a alteração de *pcond* de zero para um significará mudar de não haver possibilidade de ser condenado para ser condenado com certeza. Mesmo essa grande mudança reduz a probabilidade de ser preso em somente 0,162; aumentando *pcond* em 0,5 decresce a probabilidade de ser preso em 0,081.

O efeito das prisões é dado pelo coeficiente de *ptemp86*. Se um homem estiver na cadeia ele não pode ser preso. Como *ptemp86* é medido em meses, seis meses a mais na prisão reduzem a probabilidade de ser preso em 0,022(6) = 0,132. A equação (7.31) fornece outro exemplo no qual o modelo de probabilidade linear não pode ser verdadeiro em toda a amplitude das variáveis independentes. Se um homem esteve preso durante todos os doze meses do ano de 1986, ele não pôde ser preso nesse ano. Definindo todas as outras variáveis iguais a zero, a probabilidade prevista de prisão quando *temp86* = 12 será 0,441 − 0,022(12) = 0,177, que é diferente de zero. Mesmo assim, se partirmos da probabilidade incondicional de prisão, 0,277, doze meses na prisão reduzem a probabilidade a praticamente zero: 0,277 − 0,022(12) = 0,13.

Finalmente, o emprego reduz a probabilidade de ser preso de maneira significante. Com todos os outros fatores mantidos fixos, um homem empregado em todos os quatro trimestres tem 0,172 menos probabilidade de ser preso do que outro desempregado.

Também podemos incluir variáveis *dummy* independentes em modelos com variável *dummy* dependente. O coeficiente mede a diferença prevista na probabilidade quando a variável *dummy* vai de zero a um. Por exemplo, se adicionarmos duas *dummies* de raça, *negro* e *hispan*, à equação sobre prisões, obteremos

$$\begin{aligned} \widehat{prîs86} &= 0{,}380 - 0{,}152\,pcond + 0{,}0046\,sentmed - 0{,}0026\,temptot \\ &\qquad (0{,}019)\quad (0{,}021)\qquad (0{,}0064)\qquad (0{,}0049) \\ &\quad - 0{,}024\,ptemp86 - 0{,}038\,empr86 + 0{,}170\,negro + 0{,}096\,hispan \\ &\qquad (0{,}005)\qquad (0{,}005)\qquad (0{,}024)\qquad (0{,}021) \\ &\qquad n = 2.725,\ R^2 = 0{,}0682. \end{aligned} \tag{7.32}$$

O coeficiente de *negro* significa que, todos os outros fatores iguais, um homem negro tem uma probabilidade 0,17 maior de ser preso que um homem branco (o grupo base). Outra maneira de expressar isso é dizer que a probabilidade de prisão é 17 pontos percentuais mais elevada para os negros do que para os brancos. A diferença também é estatisticamente significante. De forma semelhante, homens hispânicos têm uma probabilidade 0,096 maior de ser presos que um homem branco.

Qual é a probabilidade de prisão prevista para um homem negro que nunca tenha sido condenado — portanto *pcond*, *sentmed*, *temptot* e *ptemp86* são todas zero — e que esteve empregado em todos os quatro trimestres de 1986? Isso parece razoável?

7.6 UM POUCO MAIS SOBRE ANÁLISE E AVALIAÇÃO DE POLÍTICAS E PROGRAMAS GOVERNAMENTAIS

Vimos alguns exemplos de modelos contendo variáveis *dummy* que podem ser úteis para a avaliação das políticas de governo. O Exemplo 7.3 deu um exemplo de avaliação de programa governamental, na qual algumas empresas receberam subsídios para treinamento de pessoal e outras não.

Como mencionamos anteriormente, precisamos ser cuidadosos ao avaliarmos programas de governo, pois na maioria dos exemplos em ciências sociais, os grupos de controle e de tratamento não são determinados aleatoriamente. Considere novamente o estudo de Holzer et al. (1993), de acordo com o qual estamos agora interessados no efeito dos subsídios para treinamento de pessoal sobre a produtividade dos trabalhadores (em contraposição à quantidade de treinamento). A equação de interesse é

$$\log(rejei) = \beta_0 + \beta_1 subs + \beta_2 \log(vendas) + \beta_3 \log(empreg) + u,$$

onde *rejei* é a taxa de rejeição dos produtos da empresa, e as últimas duas variáveis são incluídas como controles. A variável binária *subs* indica se a empresa recebeu subsídios para treinamento de pessoal em 1988.

Antes de examinarmos as estimativas, devemos nos preocupar em saber se os fatores não observados que afetam a produtividade dos trabalhadores — tais como os níveis médios de educação, aptidão, experiência e permanência no emprego — podem estar correlacionados com o fato de a empresa ter recebido subsídios. Holzer et al. destacam que os subsídios foram concedidos na ordem de chegada das solicitações. Isso não é a mesma coisa que distribuir os subsídios de forma aleatória. Pode ser que empresas com trabalhadores menos produtivos tenham visto uma oportunidade de melhorar a produtividade e, portanto, tenham sido mais diligentes na solicitação de subsídios.

Utilizando os dados contidos no arquivo JTRAIN.RAW de 1988 — quando as empresas foram efetivamente qualificadas a receber subsídios — obtemos

$$\begin{aligned} \widehat{\log(rejei)} &= 4{,}99 - 0{,}052\, subs - 0{,}455 \log(vendas) \\ &\qquad (4{,}66)\ (0{,}431) \qquad (0{,}373) \\ &\quad + 0{,}639 \log(empreg) \\ &\qquad (0{,}365) \\ &\quad n = 50,\ R^2 = 0{,}072. \end{aligned} \tag{7.33}$$

(Dezessete das cinqüenta empresas receberam subsídios para treinamento, e o índice médio de rejeição entre todas as empresas é de 3,47.) A estimativa de $-0{,}052$ de *subs* significa que, para *vendas* e *empreg* dados, as empresas que receberam subsídios têm índice de rejeição cerca de 5,2% menor que o daquelas que não receberam. Esta é a tendência do efeito esperado se os subsídios ao treinamento forem eficientes, mas a estatística t é muito pequena. Assim, desta análise de corte transversal, devemos concluir que os subsídios não tiveram efeito sobre a produtividade das empresas. Retornaremos a este exemplo no Capítulo 9 e mostraremos como a adição de informações de um ano anterior conduz a conclusões bastante diferentes.

Mesmo nos casos em que a análise das políticas de governo não envolve a atribuição de unidades para um grupo de controle e para um grupo de tratamento, devemos ser cuidadosos ao incluirmos fatores que possam estar sistematicamente relacionados com as variáveis independentes binárias de

interesse. Um bom exemplo disso é fazer um teste sobre discriminação racial. Raça não é algo determinado por um indivíduo ou por administradores governamentais. De fato, raça parece ser um exemplo perfeito de uma variável explicativa exógena, já que é determinada quando do nascimento da pessoa. Porém, por razões históricas, raça não é necessariamente exógena: existem diferenças culturais sistemáticas entre as raças, e essas diferenças podem ser importantes em um teste sobre a discriminação *corrente*.

Como um exemplo, considere a possibilidade de fazermos um teste para verificar a discriminação na aprovação de empréstimos. Se pudermos colher informações sobre, digamos, solicitações de empréstimos hipotecários individuais, poderemos definir a variável *dummy* dependente *aprovado* como igual a um se uma solicitação foi aprovada, e zero, caso contrário. Uma diferença sistemática nas taxas de aprovação entre as raças será uma indicação de discriminação. Porém, como a aprovação depende de muitos outros fatores, inclusive renda, riqueza, classificação do risco de crédito e capacidade de pagamento do empréstimo, devemos controlar esses fatores *se* neles houver diferenças sistemáticas entre as raças. Um modelo de probabilidade linear para testar a discriminação pode parecer com o seguinte:

$$aprovado = \beta_0 + \beta_1 não\text{-}branco + \beta_2 renda + \beta_3 riqueza + \beta_4 riscodecrédito + outros\ fatores,$$

A discriminação contra as minorias é indicada por uma rejeição de H_0: $\beta_1 = 0$ em favor de H_0: $\beta_1 < 0$, porque β_1 é o montante pelo qual a probabilidade de uma pessoa não-branca obter uma aprovação difere da probabilidade de uma pessoa branca obtê-la, dados os mesmos níveis das outras variáveis na equação. Se *renda*, *riqueza* etc. forem sistematicamente diferentes entre as raças, então será importante controlar esses fatores em uma análise de regressão múltipla.

Outro problema que freqüentemente surge em avaliações de políticas e programas de governo é que os indivíduos (ou empresas ou cidades) escolhem participar ou não de certos procedimentos ou programas. Por exemplo, as pessoas escolhem fazer uso de drogas ilegais ou de bebidas alcoólicas. Se quisermos examinar os efeitos de tais comportamentos sobre a situação de desemprego, a renda, ou a atividade criminal, deveremos nos preocupar com o fato de que o uso de drogas pode estar correlacionado a outros fatores que possam afetar os resultados do emprego e da criminalidade. Crianças qualificadas para participar de programas de desenvolvimento da saúde infantil são incluídas nesses programas com base na decisão de seus pais. Como a formação familiar é levada em conta nas decisões desses programas e afeta os resultados dos alunos, devemos controlar esses fatores ao verificarmos seus efeitos [veja, por exemplo, Currie e Thomas (1995)]. Indivíduos selecionados por empregadores ou por agências governamentais para participarem de programas de treinamento no emprego podem participar ou não, e essa decisão provavelmente não será aleatória [veja, por exemplo, Lynch (1991)]. Cidades e estados decidem implementar certas leis de controle de armas, e muito provavelmente essa decisão estará sistematicamente relacionada a outros fatores que afetam os crimes violentos [veja, por exemplo, Kleck e Patterson (1993)].

O parágrafo anterior dá exemplos do que, de forma geral, é conhecido como problemas de **auto-seleção** em economia. Literalmente, o termo advém do fato de que os indivíduos se auto-selecionam para certos procedimentos ou programas: a participação não é determinada de forma aleatória. O termo é geralmente usado quando um indicador binário de participação puder estar sistematicamente relacionado com fatores não-observados. Assim, se escrevermos o modelo simples

$$y = \beta_0 + \beta_1 partic + u, \qquad (7.34)$$

onde y é uma variável de resultado e *partic* é uma variável binária igual à unidade se o indivíduo, empresa ou cidade participa de um procedimento, programa, ou tem certo tipo de lei, então estamos preocupados com que o valor médio de u dependa da participação: $E(u|partic = 1) \neq E(u|partic = 0)$. Como sabemos, isso levará o estimador β_1 da regressão simples a ser viesado, e portanto não encontraremos o efeito verdadeiro da participação. Assim o problema da auto-seleção é outra maneira de uma variável explicativa (*partic*, neste caso) poder ser endógena.

Por enquanto sabemos que a análise de regressão múltipla pode, até certo ponto, aliviar o problema de auto-seleção. Fatores no termo erro na equação (7.34) que sejam correlacionados com *partic* podem ser incluídos em uma equação de regressão múltipla, assumindo, é claro, que possamos coletar os dados desses fatores. Infelizmente, em muitos casos, preocupa-nos que fatores não-observados estejam relacionados com a participação, caso em que a regressão múltipla produzirá estimadores viesados.

Na análise padrão de regressão múltipla que usa dados de corte transversal, devemos ficar atentos com o aparecimento de efeitos espúrios de programas nas variáveis de resultado devido ao problema de auto-seleção. Um bom exemplo disto está contido em Currie e Cole (1993). Esses autores examinam os efeitos da participação, em um programa específico de auxílio para famílias com dependentes menores de idade, sobre o peso dos recém-nascidos. Mesmo após controlar uma variedade de famílias e características culturais, os autores obtêm estimativas MQO que sugerem que a participação nesse programa específico *diminui* o peso dos recém-nascidos. Como os autores ressaltam, é difícil acreditar que a participação no programa, por si própria, seja a *causa* do peso menor dos recém-nascidos. [Veja Currie (1995) para exemplos adicionais.] Utilizando um método econométrico diferente, que discutiremos no Capítulo 15, Currie e Cole encontraram evidência tanto para nenhum efeito como para um efeito positivo da participação no programa de auxílio para famílias com dependentes menores de idade, sobre o peso dos recém-nascidos.

Quando o problema de auto-seleção faz com que a análise de regressão múltipla seja viesada devido à falta de suficientes variáveis de controle, métodos mais avançados, tratados nos Capítulos 13, 14 e 15 poderão ser usados.

Neste capítulo aprendemos como usar informações qualitativas na análise de regressão. No caso mais simples, uma variável *dummy* é definida para fazer a distinção entre dois grupos, e o coeficiente estimado da variável *dummy* estima a diferença *ceteris paribus* entre os dois grupos. A admissão de mais de dois grupos é realizada pela definição de um conjunto de variáveis *dummy*: se houver g grupos, então $g - 1$ variáveis *dummy* são incluídas no modelo. Todas as estimativas das variáveis *dummy* são interpretadas em relação ao grupo base ou referencial (o grupo para o qual nenhuma variável *dummy* é incluída no modelo).

As variáveis *dummy* também são úteis para incorporar informações ordinais, como classificações de crédito e de aparência pessoal, em modelos de regressão. Simplesmente definimos um conjunto de variáveis *dummy* representando os diferentes resultados da variável ordinal, admitindo uma das categorias como grupo base.

Para possibilitar diferenças de inclinações entre os diferentes grupos, as variáveis *dummy* podem interagir com variáveis quantitativas. No caso extremo, podemos permitir que cada grupo tenha sua própria inclinação em todas as variáveis, como também seu próprio intercepto. O teste de Chow pode ser usado para detectar se existem quaisquer diferenças entre os grupos. Em muitos casos, é mais interessante verificar se, após termos permitido uma diferença de interceptos, as inclinações de dois gru-

pos diferentes são as mesmas. Um teste F padrão pode ser usado para esse propósito em um modelo irrestrito que inclua interações entre a *dummy* do grupo e todas as variáveis.

O modelo de probabilidade linear, que é simplesmente estimado pelo MQO, possibilita explicar uma resposta binária usando análise de regressão. As estimativas MQO agora são interpretadas como alterações na probabilidade de "sucesso" ($y = 1$), dado um aumento de uma unidade na variável explicativa correspondente. O MPL tem algumas inconveniências: pode produzir probabilidades previstas menores que zero ou maiores que um, implica um efeito marginal constante de cada variável explicativa que aparece em sua forma original, e contém heteroscedasticidade. Os primeiros dois problemas muitas vezes não são graves quando estamos obtendo estimativas dos efeitos parciais das variáveis explicativas na faixa intermediária dos dados. A heteroscedasticidade invalida os erros-padrão usuais do MQO e as estatísticas de testes mas, como veremos no próximo capítulo, isso é facilmente corrigido em amostras suficientemente grandes.

Terminamos este capítulo com uma discussão de como variáveis binárias são usadas para a avaliação de políticas e programas de governo. Como em toda análise de regressão, devemos nos lembrar que a participação em programas, ou algum outro regressor binário com implicações relacionadas a políticas governamentais, pode ser correlacionada com fatores não observados que afetem a variável dependente, resultando no viés usual de variáveis omitidas.

7.1 Utilizando os dados contidos no arquivo SLEEP75.RAW (veja também o Problema 3.3), obtemos a equação estimada

$$\widehat{dormir} = 3.840{,}83 - 0{,}163\ trabtot - 11{,}71\ educ - 8{,}70\ idade$$
$$(235{,}11)\quad (0{,}018)\qquad\quad (5{,}86)\qquad (11{,}21)$$
$$+ 0{,}128\ idade^2 + 87{,}75\ masculino$$
$$(0{,}134)\qquad\quad (34{,}33)$$
$$n = 706,\ R^2 = 0{,}123,\ \overline{R}^2 = 0{,}117.$$

A variável *dormir* é o total de minutos gastos por semana dormindo durante a noite, *trabtot* é o total de minutos semanais gastos trabalhando, *educ* e *idade* são medidas em anos e *masculino* é uma variável *dummy* de gênero.

(i) Supondo todos os outros fatores iguais, existe evidência de que os homens durmam mais que as mulheres? O quanto essa evidência é forte?

(ii) Existe uma relação de substituição estatisticamente significante entre trabalhar e dormir? Qual é a relação de substituição estimada?

(iii) Que outras regressões você precisa executar para testar a hipótese nula de que, mantendo fixos os outros fatores, a idade não tem efeito sobre dormir?

7.2 As seguintes equações foram estimadas utilizando os dados contidos no arquivo BWGHT.RAW:

$$\log(\textit{pesonas}) = 4{,}66 - 0{,}0044\ \textit{cigs} + 0{,}0093\ \log(\textit{rendfam}) + 0{,}016\ \textit{ordnas}$$
$$(0{,}22)\quad (0{,}0009)\qquad (0{,}0059)\qquad\qquad (0{,}006)$$
$$+ 0{,}027\ \textit{masculino} + 0{,}055\ \textit{branco}$$
$$(0{,}010)\qquad\qquad (0{,}013)$$
$$n = 1.388,\ R^2 = 0{,}0472$$

e

$$\log(\textit{pesonas}) = 4{,}65 - 0{,}0052\ \textit{cigs} + 0{,}0110\ \log(\textit{rendfam}) + 0{,}017\ \textit{ordnas}$$
$$(0{,}38)\quad (0{,}0010)\qquad (0{,}0085)\qquad\qquad (0{,}006)$$
$$+ 0{,}034\ \textit{masculino} + 0{,}045\ \textit{branco} - 0{,}0030\ \textit{educm} + 0{,}0032\ \textit{educp}$$
$$(0{,}011)\qquad\qquad (0{,}015)\qquad\qquad (0{,}0030)\qquad\qquad (0{,}0026)$$
$$n = 1.191,\ R^2 = 0{,}0493.$$

As variáveis são definidas como no Exemplo 4.9, mas adicionamos uma variável *dummy* para o caso de a criança ser do sexo masculino e uma variável *dummy* indicando se a criança é classificada como branca.

(i) Na primeira equação, interprete o coeficiente da variável *cigs*. Particularmente, qual é o efeito no peso dos recém-nascidos se a mãe fumar dez ou mais cigarros por dia?

(ii) Quanto se espera que um recém-nascido branco pesará mais que uma criança não-branca, mantendo fixos todos os outros fatores na primeira equação? A diferença é estatisticamente significante?

(iii) Comente sobre o efeito estimado e a significância estatística de *educm*.

(iv) Com a informação dada, por que você não terá condições de computar a estatística F da significância conjunta de *educm* e *educp*? O que você teria que fazer para computar a estatística F?

7.3 Utilizando os dados contidos no arquivo GPA2.RAW, a seguinte equação foi estimada:

$$\textit{sât} = 1.028{,}10 + 19{,}30\ \textit{tamclas} - 2{,}19\ \textit{tamclas}^2 - 45{,}09\ \textit{feminino}$$
$$(6{,}29)\quad (3{,}83)\qquad\qquad (0{,}53)\qquad\qquad (4{,}29)$$
$$- 169{,}81\ \textit{negro} + 62{,}31\ \textit{feminino} \cdot \textit{negro}$$
$$(12{,}71)\qquad\qquad (18{,}15)$$
$$n = 4.137,\ R^2 = 0{,}0858.$$

A variável *sat* é a nota combinada de matemática e habilidade verbal do estudante para ingresso em curso superior, *tamclas* é o tamanho da classe do aluno no ensino médio, em centenas, *feminino* é uma variável *dummy* para gênero e *negro* é uma variável *dummy* da raça igual a um para negros e zero, caso contrário.

(i) Existe evidência forte que $tamclas^2$ deveria ser incluída no modelo? Desta equação, qual é o tamanho ótimo da classe no ensino médio?

(ii) Mantendo fixo *tamclas*, qual é a diferença estimada na nota *sat* entre mulheres não-negras e homens não-negros? O quanto é estatisticamente significante essa diferença estimada?

(iii) Qual é a diferença estimada na *sat* entre homens não-negros e homens negros? Teste a hipótese nula de que não há diferença entre suas notas, contra a hipótese alternativa de que existe uma diferença.

(iv) Qual é a diferença estimada na nota *sat* entre mulheres negras e mulheres não-negras? O que você necessitaria fazer para verificar se a diferença é estatisticamente significante?

7.4 Uma equação que explica os salários dos diretores executivos é

$$\begin{aligned}\log(\textit{sâlário}) = {} & 4{,}59 + 0{,}257\log(\textit{vendas}) + 0{,}011\,\textit{rma} + 0{,}158\,\textit{financeira} \\ & \quad (0{,}30)\quad (0{,}032)\qquad\qquad (0{,}004)\qquad (0{,}089) \\ & + 0{,}181\,\textit{prodcons} - 0{,}283\,\textit{serv} \\ & \quad (0{,}085)\qquad\quad (0{,}099) \\ & n = 209,\ R^2 = 0{,}357.\end{aligned}$$

Os dados usados estão no arquivo CEOSAL1.RAW, no qual *financeira*, *prodcons* e *serv* são variáveis binárias indicando as empresas financeiras, de produtos de consumo e de serviços públicos. O ramo de atividade omitido foi o de transportes.

(i) Calcule a diferença percentual aproximada no salário estimado entre os setores de serviços públicos e de transportes, mantendo fixos *vendas* e *rma*. A diferença é estatisticamente significante ao nível de 1%?

(ii) Use a equação (7.10) para obter a diferença percentual exata no salário estimado entre os setores de serviços públicos e de transportes e compare-a com a resposta obtida na parte (i).

(iii) Qual é a diferença percentual aproximada no salário estimado entre as indústrias de produtos de consumo e financeiros? Escreva uma equação que possibilite verificar se a diferença é estatisticamente significante.

7.5 No Exemplo 7.2 defina *semPC* como uma variável *dummy* igual a um se o aluno não possuir um PC, e zero caso contrário.

(i) Se *semPC* for usada no lugar de *PC* na equação (7.6), o que acontece com o intercepto na equação estimada? Qual será o coeficiente de *semPC*? (*Sugestão*: Escreva $PC = 1 - semPC$ e agregue isso na equação $\hat{nmgrad} = \hat{\beta}_0 + \hat{\delta}_0 PC + \hat{\beta}_1 nmem + \hat{\beta}_2 tac$.)

(ii) O que acontecerá com o R-quadrado se *semPC* for usado em lugar de *PC*?

(iii) As variáveis *PC* e *semPC* deveriam ser incluídas como variáveis independentes no modelo? Explique.

7.6 Para testar a eficiência de um programa de treinamento de pessoal sobre os subseqüentes salários dos trabalhadores, especificamos o modelo

$$\log(\textit{salário}) = \beta_0 + \beta_1 \textit{trein} + \beta_2 \textit{educ} + \beta_3 \textit{exper} + u,$$

onde *trein* é uma variável binária igual à unidade se um trabalhador participou do programa. Pense no termo erro u como contendo a aptidão não observada do trabalhador. Se trabalhadores menos aptos tiverem maior oportunidade de serem selecionados para o programa, e você usar uma análise MQO, o que você pode dizer sobre o provável viés no estimador MQO de β_1? (*Sugestão*: Consulte o Capítulo 3.)

7.7 No exemplo na equação (7.29), suponha que definamos *foraft* como sendo um se a mulher estiver fora da força de trabalho, e zero caso contrário.

(i) Se fizermos a regressão de *foraft* sobre todas as variáveis independentes na equação (7.29), o que acontecerá com as estimativas do intercepto e da inclinação? (Sugestão: $naft = 1 - foraft$. Agregue essa expressão na equação populacional $naft = \beta_0 + \beta_1 nesprend + \beta_2 educ + ...$ e reorganize.)

(ii) O que acontecerá com os erros-padrão das estimativas do intercepto e da inclinação?

(iii) O que acontecerá com o R-quadrado?

7.8 Suponha que você colete dados de uma pesquisa sobre salários, educação, experiência e gênero. Além disso, você solicita informações sobre o uso de maconha. A pergunta original é: "Em quantas ocasiões distintas, no mês passado, você fumou maconha?".

(i) Escreva uma equação que permita a você estimar os efeitos do uso de maconha sobre os salários com todos os outros fatores controlados. Você deve ter condições de fazer declarações do tipo, "Estima-se que fumar maconha cinco vezes ou mais por mês altera os salários em $x\%$."

(ii) Escreva um modelo que permita verificar se o uso de drogas tem efeitos diferentes sobre os salários dos homens e das mulheres. Como você verificaria que não existem diferenças nos efeitos do uso de drogas nos homens e nas mulheres?

(iii) Suponha que você considere ser melhor avaliar o uso de maconha colocando as pessoas em uma de quatro categorias: não-usuário, usuário leve (um a cinco vezes por mês), usuário moderado (seis a dez vezes por mês) e usuário inveterado (mais de dez vezes por mês). Agora escreva um modelo que permita estimar os efeitos da maconha sobre os salários.

(iv) Usando o modelo da parte (iii), explique em detalhes como testar a hipótese nula de que o uso de maconha não tem efeito sobre o salário. Seja bastante específico e inclua uma relação cuidadosa de graus de liberdade.

(v) Quais são alguns dos problemas potenciais de procurar inferência causal utilizando os dados da pesquisa que você coletou?

Heteroscedasticidade

hipótese de homoscedasticidade, apresentada no Capítulo 3 para a regressão múltipla, significa que a variância do erro não observável, u, condicional nas variáveis explicativas, é constante. A homoscedasticidade não se mantém sempre que a variância dos fatores não-observáveis muda ao longo de diferentes segmentos da população, nos quais os segmentos são determinados pelos diferentes valores das variáveis explicativas. Por exemplo, em uma equação de poupança, a heteroscedasticidade está presente se a variância dos fatores não-observados que afetam a poupança aumenta com a renda.

Nos Capítulos 4 e 5, vimos que a homoscedasticidade é necessária para justificar os habituais testes t e F, bem como os intervalos de confiança da estimação MQO do modelo de regressão linear, mesmo com amostras de tamanhos grandes. Neste capítulo discutiremos as soluções disponíveis quando ocorre heteroscedasticidade, e também mostraremos como verificar sua presença. Iniciamos fazendo uma breve revisão das conseqüências da heteroscedasticidade para a estimação de mínimos quadrados ordinários.

8.1 CONSEQÜÊNCIAS DA HETEROSCEDASTICIDADE PARA O MÉTODO MQO

Considere novamente o modelo de regressão linear múltipla:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u. \qquad \textbf{(8.1)}$$

No Capítulo 3, provamos a inexistência de viés dos estimadores de MQO $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 \ldots, \hat{\beta}_k$ sob as quatro primeiras hipóteses de Gauss-Markov, RLM.1 a RLM.4. No Capítulo 5 mostramos que as mesmas quatro hipóteses implicam consistência dos estimadores de MQO. A hipótese de homoscedasticidade RLM.5, estabelecida em termos da variância do erro como $\text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$, não teve participação para mostrar se os estimadores MQO eram não-viesados ou consistentes. É importante lembrar que a heteroscedasticidade não provoca viés ou inconsistência nos estimadores MQO de β_j, enquanto algo como a omissão de uma variável importante teria esse efeito.

A interpretação de nossas medidas dos graus-de-ajuste, R^2 e $\overline{R}^2$, também não é afetada pela presença de heteroscedasticidade. Por quê? Lembre-se de que na Seção 6.3 o R-quadrado usual e o R-quadrado ajustado são modos diferentes de estimar o R-quadrado da população, que é simplesmente $1 - \sigma_u^2/\sigma_y^2$, onde σ_u^2 é a variância do erro da população e σ_y^2 é a variância populacional de y. A questão crucial é

que, como ambas as variâncias no R-quadrado da população são variâncias incondicionais, o R-quadrado da população não é afetado pela presença de heteroscedasticidade em $\text{Var}(u|x_1, ..., x_k)$. Além disso, SQR/$n$ estima consistentemente σ_u^2, e SQT/n estima consistentemente σ_y^2, seja $\text{Var}(u|x_1, ..., x_k)$ constante ou não. O mesmo é verdadeiro quando usamos os ajustes dos graus de liberdade. Portanto, R^2 e $\bar{R}^2$ são ambos estimadores consistentes do R-quadrado da população mantendo-se ou não a hipótese de homoscedasticidade.

Se a heteroscedasticidade não provoca viés ou inconsistência nos estimadores MQO, por que a introduzimos como uma das hipóteses de Gauss-Markov? Lembre-se do Capítulo 3 que os estimadores de *variâncias*, $\text{Var}(\hat{\beta}_j)$, são viesados sem a hipótese de homoscedasticidade. Como os erros-padrão dos estimadores MQO são baseados diretamente nessas variâncias, eles não mais são válidos para construirmos intervalos de confiança e estatísticas t. As estatísticas t habituais dos estimadores MQO não têm distribuições t na presença de heteroscedasticidade, e o problema não será resolvido com o uso de amostras de tamanho grande. Veremos isso claramente no caso de regressão simples na próxima seção, na qual derivaremos a variância do estimador MQO da inclinação sob heteroscedasticidade e proporemos um estimador válido na presença de heteroscedasticidade. De maneira semelhante, as estatísticas F não têm distribuição F, e a estatística LM não tem uma distribuição qui-quadrada assimptótica. Em resumo, as estatísticas que usamos para testar hipóteses sob as hipóteses de Gauss-Markov não são válidas na presença de heteroscedasticidade.

Também sabemos que o teorema de Gauss-Markov, que diz que os estimadores MQO são os melhores estimadores lineares não-viesados, vale-se de forma crucial da hipótese de homoscedasticidade. Se $\text{Var}(u|\boldsymbol{x})$ não for constante, os estimadores MQO não mais serão BLUE. Além disso, os estimadores MQO não mais serão assimptoticamente eficientes na classe dos estimadores descritos no Teorema 5.3. Como veremos na Seção 8.4, é possível encontrar estimadores que são mais eficientes que os MQO na presença de heteroscedasticidade (embora seja necessário o conhecimento da forma da heteroscedasticidade). Com amostras de tamanhos relativamente grandes, pode não ser tão importante obter um estimador eficiente. Na próxima seção mostraremos como os testes estatísticos usuais dos estimadores MQO podem ser modificados de forma a serem válidos, pelo menos assimptoticamente.

8.2 INFERÊNCIA ROBUSTA EM RELAÇÃO À HETEROSCEDASTICIDADE APÓS A ESTIMAÇÃO MQO

Como os testes de hipóteses são um componente importante de qualquer análise econométrica e a inferência habitual dos estimadores MQO geralmente é imperfeita na presença de heteroscedasticidade, temos que decidir se abandonaremos de vez o método MQO. Felizmente, ele ainda é útil. Nas últimas duas décadas, os econometristas aprenderam como ajustar erros-padrão, estatísticas t, F e LM de forma a torná-las válidas na presença de **heteroscedasticidade de forma desconhecida**. Isto é muito conveniente, pois significa que podemos descrever novas estatísticas que funcionam independentemente do tipo de heteroscedasticidade presente na população. Os métodos desta seção são conhecidos como procedimentos *robustos em relação à heteroscedasticidade*, porque eles são válidos — pelo menos em amostras grandes — tenham ou não os erros variância constante, e não precisamos saber qual é o caso.

Começamos esboçando como as variâncias, $\text{Var}(\hat{\beta}_j)$, podem ser estimadas na presença de heteroscedasticidade. Uma derivação cuidadosa da teoria está bem além do escopo desta obra, mas a aplicação de métodos robustos em relação à heteroscedasticidade é bastante fácil, pois muitos programas estatísticos e econométricos computam essas estatísticas como uma opção.

Primeiro, considere o modelo com uma única variável independente, na qual incluímos um subscrito i por ênfase:

$$y_i = \beta_0 + \beta_1 x_i + u_i.$$

Assumimos que as primeiras quatro hipóteses de Gauss-Markov se sustentam. Se os erros contiverem heteroscedasticidade, então

$$\text{Var}(u_i|x_i) = \sigma_i^2,$$

onde colocamos um subscrito i em σ^2 para indicar que a variância do erro depende do valor particular de x_i.

Escreva o estimador MQO como

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x})\, u_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Sob as hipóteses RLM.1 a RLM.4 (isto é, sem a hipótese de homoscedasticidade), e condicionado aos valores de x_i na amostra, podemos usar os mesmos argumentos do Capítulo 2 para mostrar que

$$\text{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sigma_i^2}{\text{SQT}_x^2}, \tag{8.2}$$

onde $\text{SQT}_x = \sum_{i=1}^{n} (x_i - \bar{x})^2$ é a soma dos quadrados total de x_i. Quando $\sigma_i^2 = \sigma^2$ para todo i, essa fórmula se reduz à forma habitual, σ^2/SQT_x. A equação (8.2) mostra explicitamente que, no caso da regressão simples, a fórmula de variância derivada sob homoscedasticidade não mais é válida quando a heteroscedasticidade está presente.

Como o erro-padrão de $\hat{\beta}_1$ é baseado diretamente na estimativa de $\text{Var}(\hat{\beta}_1)$, precisamos de um modo de estimar a equação (8.2) quando a heteroscedasticidade está presente. White (1980) mostrou como isso pode ser feito. Façamos $\hat{u}_i$ representar os resíduos MQO da regressão inicial de y sobre x. Então, um estimador válido de $\text{Var}(\hat{\beta}_1)$, para a heteroscedasticidade de qualquer forma (inclusive homoscedasticidade), é

$$\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \hat{u}_i^2}{\text{SQT}_x^2}, \tag{8.3}$$

que é facilmente calculado após a regressão MQO.

Em que sentido a expressão (8.3) é um estimador válido de $\text{Var}(\hat{\beta}_1)$? Isso é bastante sutil. Resumidamente, pode ser mostrado que quando a equação (8.3) é multiplicada pelo tamanho da amostra, n, ela converge em probabilidade para $\text{E}[(x_i - \mu_x)^2 u_i^2]/(\sigma_x^2)^2$, que é o limite de probabilidade de n vezes (8.2). Em última análise, isso é o necessário para justificar o uso de erros-padrão para construir

intervalos de confiança e estatísticas *t*. A lei dos grandes números e o teorema do limite central desempenham papéis importantes no estabelecimento dessas convergências. Você pode consultar o artigo original de White para detalhes, mas ele é bastante técnico. Veja também Wooldridge (2002, Capítulo 4).

Uma fórmula semelhante funciona no modelo geral de regressão múltipla

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u,$$

Pode ser mostrado que um estimador válido de $\text{Var}(\hat{\beta}_j)$, sob as hipóteses RLM.1 a RLM.4, é

$$\widehat{\text{Var}}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{\text{SQR}_j^2}, \qquad \textbf{(8.4)}$$

onde $\hat{r}_{ij}$ representa o i-ésimo resíduo da regressão de x_j sobre todas as outras variáveis independentes, e SQR_j é a soma dos resíduos quadrados dessa regressão (veja Seção 3.2 para a representação parcial das estimativas MQO). A raiz quadrada de (8.4) é chamada de **erro-padrão robusto em relação à heteroscedasticidade** de $\hat{\beta}_j$. Em econometria, esses erros-padrão robustos são em geral atribuídos a White (1980). Trabalhos anteriores em estatística, notavelmente os de Eicker (1967) e Huber (1967), indicaram a possibilidade de obter tais erros-padrão robustos. Em trabalhos aplicados, eles são algumas vezes chamados de *erros-padrão de White, de Huber*, ou *de Eicker* (ou alguma combinação hifenizada desses nomes). Faremos referência a eles apenas como erros-padrão robustos em relação à heteroscedasticidade, ou mesmo apenas como erros-padrão robustos, quando o contexto for claro.

Algumas vezes, como uma correção de graus de liberdade, a equação (8.4) é multiplicada por $n/(n-k-1)$ antes de extrairmos a raiz quadrada. O raciocínio para esse ajuste é que, se os resíduos quadrados MQO $\hat{u}_i^2$ fossem os mesmos para todas as observações i — a forma mais forte possível de homoscedasticidade em uma amostra — obteríamos os erros-padrão habituais MQO. Outras modificações de (8.4) são estudadas em MacKinnon e White (1985). Como todas as formas têm apenas justificativas assimptóticas e são assimptoticamente equivalentes, nenhuma delas é uniformemente preferida às outras. Em geral, utilizamos qualquer forma que seja computada pelo programa econométrico em uso.

Uma vez que os erros-padrão robustos em relação à heteroscedasticidade tenham sido obtidos, é fácil construir uma **estatística *t* robusta em relação à heteroscedasticidade**. Lembre-se de que a forma geral da estatística *t* é

$$t = \frac{\textit{estimativa} - \textit{valor hipotético}}{\textit{erro-padrão}}. \qquad \textbf{(8.5)}$$

Como ainda estamos usando as estimativas MQO e escolhemos o valor hipotético antecipadamente, a única diferença entre a estatística *t* usual de MQO e a estatística *t* robusta em relação à heteroscedasticidade é como o erro-padrão é calculado.

EXEMPLO 8.1

(Equação do Log dos Salários com Erros-Padrão Robustos em Relação à Heteroscedasticidade)

Estimamos o modelo no Exemplo 7.6, mas escrevemos os erros-padrão robustos em relação à heteroscedasticidade juntamente com os erros-padrão do MQO. Algumas das estimativas estão registradas com mais dígitos para podermos comparar os erros-padrão habituais com os erros-padrão robustos em relação à heteroscedasticidade:

$$\begin{aligned}\widehat{\log(salário)} = {} & 0{,}321 + 0{,}213\,hcasados - 0{,}198\,mcasadas - 0{,}110\,msolteiras \\ & \quad (0{,}100)\ (0{,}055) \qquad (0{,}058) \qquad (0{,}056) \\ & \quad [0{,}109]\ [0{,}057] \qquad [0{,}058] \qquad [0{,}057] \\ & + 0{,}0789\,educ + 0{,}0268\,exper - 0{,}00054\,exper^2 \\ & \quad (0{,}0067) \qquad (0{,}0055) \qquad (0{,}00011) \\ & \quad [0{,}0074] \qquad [0{,}0051] \qquad [0{,}00011] \\ & + 0{,}0291\,perm - 0{,}00053\,perm^2 \\ & \quad (0{,}0068) \qquad (0{,}00023) \\ & \quad [0{,}0069] \qquad [0{,}00024] \\ & n = 526,\ R^2 = 0{,}461.\end{aligned} \tag{8.6}$$

Os erros-padrão usuais de MQO estão entre parênteses, ( ), abaixo das estimativas MQO correspondentes, e os erros-padrão robustos em relação à heteroscedasticidade estão entre colchetes, [ ]. Os números entre colchetes são as únicas novidades, já que a equação ainda é estimada por MQO.

Diversos fatores são aparentes a partir da equação (8.6). Primeiro, nesta aplicação particular, qualquer variável que era estatisticamente significante com o uso da estatística *t* habitual, continua estatisticamente significante com o uso da estatística *t* robusta em relação à heteroscedasticidade. Isto é devido ao fato de os dois conjuntos de erros-padrão não serem muito diferentes. (Os *p*-valores associados serão levemente diferentes devido ao fato de as estatísticas *t* robustas não serem idênticas às estatísticas *t* habituais, não robustas.) A maior mudança relativa nos erros-padrão está no coeficiente de *educ*: o erro-padrão usual é 0,0067, e o erro-padrão robusto é 0,0074. Não obstante, o erro-padrão robusto implica uma estatística *t* robusta acima de 10.

A equação (8.6) também mostra que os erros-padrão robustos podem ser maiores, ou menores, do que os erros-padrão usuais. Por exemplo, o erro-padrão robusto da variável *exper* é 0,0051, enquanto o erro-padrão usual é 0,0055. Não sabemos qual será maior antecipadamente. Como um tópico empírico, os erros-padrão robustos são freqüentemente maiores do que os erros-padrão usuais.

Antes de sairmos deste exemplo, devemos enfatizar que não sabemos, neste ponto, sequer se a heteroscedasticidade está presente no modelo populacional básico da equação (8.6). Tudo que fizemos foi descrever, juntamente com os erros-padrão usuais, aqueles que são válidos (assimptoticamente) haja ou não presença de heteroscedasticidade. Podemos ver que nenhuma conclusão importante é destruída pelo uso dos erros-padrão robustos neste exemplo. Isso acontece com freqüência em trabalhos aplicados, mas, em outros casos, as diferenças entre os erros-padrão usuais e os robustos são muito maiores. Para um exemplo no qual as diferenças são substanciais, veja o Problema 8.7.

Neste ponto, você deve estar se perguntando: se os erros-padrão robustos em relação à heteroscedasticidade são válidos com maior freqüência que os erros-padrão usuais MQO, por que nos preocuparmos com os erros-padrão usuais, afinal? Está é uma pergunta sensata. Uma das razões para eles

ainda serem usados em trabalhos de corte transversal é que, se a hipótese de homoscedasticidade se mantiver e os erros forem normalmente distribuídos, as estatísticas *t* usuais têm distribuições *t* exatas, independentemente do tamanho da amostra (veja Capítulo 4). Os erros-padrão robustos e as estatísticas *t* robustas são justificadas somente quando o tamanho da amostra se torna grande. Com amostras de tamanho pequeno, as estatísticas *t* robustas podem ter distribuições que não sejam muito próximas da distribuição *t*, e isso pode ofuscar nossa inferência.

Em amostras de tamanho grande podemos tomar a decisão de sempre levar em conta somente os erros-padrão robustos em relação à heteroscedasticidade em aplicações de corte transversal, e esta prática vem sendo seguida cada vez mais em trabalhos aplicados. Também é comum usar ambos os erros-padrão, como na equação (8.6), de maneira que um leitor possa determinar se alguma conclusão é sensível ao erro-padrão em uso.

Também é possível obter estatísticas *F* e *LM* que sejam robustas em relação à heteroscedasticidade de forma desconhecida, e mesmo arbitrária. A **estatística *F* robusta em relação à heteroscedasticidade** (ou uma transformação simples dela) também é chamada de *estatística de Wald robusta em relação à heteroscedasticidade*. Uma abordagem geral da estatística de Wald requer álgebra matricial e está esboçada no Apêndice E, disponível na página do livro, no site www.thomsonlearning.com.br; veja Wooldridge (2002, Capítulo 4) para um tratamento mais detalhado. No entanto, a utilização de estatísticas robustas em relação à heteroscedasticidade para restrições de exclusões múltiplas é simples, pois muitos programas econométricos computam essas estatísticas rotineiramente.

EXEMPLO 8.2

(Estatística *F* Robusta em Relação à Heteroscedasticidade)

Utilizando os dados *s* contidos no arquivo GPA3.RAW, estimamos a seguinte equação:

$$\begin{aligned} nmgradac = {} & 1{,}47 + 0{,}00114\,sat - 0{,}00857\,emperc + 0{,}00250\,tothrs \\ & \quad (0{,}23)\ (0{,}00018) \quad (0{,}00124) \quad (0{,}00073) \\ & \quad [0{,}22]\ [0{,}00019] \quad [0{,}00140] \quad [0{,}00073] \\ & + 0{,}303\,feminino - 0{,}128\,negro - 0{,}059\,branco \\ & \quad (0{,}059) \quad (0{,}147) \quad (0{,}141) \\ & \quad [0{,}059] \quad [0{,}118] \quad [0{,}110] \\ & n = 366,\ R^2 = 0{,}4006,\ \bar{R}^2 = 0{,}3905. \end{aligned} \tag{8.7}$$

Novamente, as diferenças entre os erros-padrão usuais e os erros-padrão robustos em relação à heteroscedasticidade não são muito grandes e o uso da estatística *t* robusta não altera a significância estatística de qualquer variável independente. Tampouco os testes de significância conjunta são muito afetados. Suponha que queiramos testar a hipótese nula de que, depois de termos todos os outros fatores controlados, não haja diferenças em *nmgradac* por raça. Isso é escrito como H_0: $\beta_{negro} = 0$, $\beta_{branco} = 0$. A estatística *F* habitual é facilmente obtida, uma vez que tenhamos o *R*-quadrado do modelo restrito; o cálculo resulta em 0,3983. A estatística *F* então é $[(0{,}4006 - 0{,}3983)/(1 - 0{,}4006)](359/2) \approx 0{,}69$. Se houver presença de heteroscedasticidade, essa versão do teste não é válida. A versão robusta em relação à heteroscedasticidade não tem uma forma simples, mas pode ser computada utilizando-se certos programas estatísticos. O valor da estatística *F* robusta em relação à heteroscedasticidade é de 0,75, diferenciando-se apenas levemente do valor da versão não robusta. O *p*-valor do teste robusto é 0,474, que não está próximo dos níveis padrão de significância. Não é possível rejeitar a hipótese nula usando qualquer um dos testes.

Computando Testes *LM* Robustos em Relação à Heteroscedasticidade

Nem todos os programas econométricos calculam estatísticas F que sejam robustas em relação à heteroscedasticidade. Portanto, algumas vezes é conveniente ter um meio de obter um teste de restrições de exclusões múltiplas que seja robusto em relação à heteroscedasticidade e não exija um tipo especial de programa econométrico. Uma **estatística *LM* robusta em relação à heteroscedasticidade** é facilmente obtida usando qualquer programa econométrico.

Avalie a seguinte declaração: Os erros-padrão robustos em relação à heteroscedasticidade são sempre maiores que os erros-padrão usuais.

Para ilustrar o cálculo da estatística *LM* robusta, considere o modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u,$$

e suponha que queiramos testar H_0: $\beta_4 = 0, \beta_5 = 0$. Para obter a estatística *LM* usual, primeiro estimamos o modelo restrito (isto é, o modelo sem x_4 e x_5) para obter os resíduos, $\tilde{u}$. Então fazemos a regressão de $\tilde{u}$ sobre todas as variáveis independentes e a estatística $LM = n \cdot R_{\tilde{u}}^2$, onde $R_{\tilde{u}}^2$ é o R-quadrado usual desta regressão.

Obter uma versão que seja robusta em relação à heteroscedasticidade exige mais trabalho. Um método de calcular a estatística requer somente regressões MQO. Precisamos dos resíduos, digamos $\tilde{r}_1$, da regressão de x_4 sobre x_1, x_2, x_3. Também necessitamos dos resíduos, digamos $\tilde{r}_2$, da regressão de x_5 sobre x_1, x_2, x_3. Assim, fazemos a regressão de cada uma das variáveis independentes excluídas, conforme a hipótese nula, sobre todas as variáveis independentes incluídas. A cada vez, guardamos os resíduos. O último passo parece estranho, mas ele é, no final das contas, apenas um artifício de cálculo. Compute a regressão de

$$1 \text{ sobre } \tilde{r}_1 \tilde{u}, \tilde{r}_2 \tilde{u}, \qquad \textbf{(8.8)}$$

sem um intercepto. Sim, na verdade definimos uma variável dependente com valor um para todas as observações. Fazemos a regressão dessa constante sobre os produtos $\tilde{r}_1 \tilde{u}$ e $\tilde{r}_2 \tilde{u}$. A estatística *LM* robusta acaba sendo $n - \text{SQR}_1$, onde SQR_1 é exatamente a soma dos resíduos quadrados usual da regressão (8.8).

A razão pela qual isso funciona é algo técnica. Basicamente, isso significa fazer para o teste *LM* o que os erros-padrão robustos fazem para o teste t. [Veja Wooldridge (1991b) ou Davidson e MacKinnon (1993) para uma discussão mais detalhada.]

Agora resumimos o cálculo da estatística *LM* robusta em relação à heteroscedasticidade no caso geral.

UMA ESTATÍSTICA *LM* ROBUSTA EM RELAÇÃO À HETEROSCEDASTICIDADE:

1. Obtenha os resíduos $\tilde{u}$ do modelo restrito.

2. Faça a regressão de cada uma das variáveis independentes excluídas, conforme a hipótese nula, sobre todas as variáveis independentes incluídas; se houver q variáveis excluídas, isso levará a q conjuntos de resíduos $(\tilde{r}_1, \tilde{r}_2, ..., \tilde{r}_q)$.
3. Encontre os produtos de cada $\tilde{r}_j$ por $\tilde{u}$ (para todas as observações).
4. Faça a regressão de 1 sobre $\tilde{r}_1\tilde{u}, \tilde{r}_2\tilde{u}, ..., \tilde{r}_q\tilde{u}$, sem um intercepto. A estatística *LM* robusta é $n - \text{SQR}_1$, onde SQR_1 é exatamente a soma dos resíduos quadrados desta última regressão. Sob H_0, a estatística *LM* é distribuída aproximadamente como χ^2_q.

Uma vez que a estatística *LM* robusta tenha sido obtida, a regra de rejeição e a computação dos *p*-valores são as mesmas da estatística *LM* usual da Seção 5.2.

EXEMPLO 8.3

(Estatística *LM* Robusta em Relação à Heteroscedasticidade)

Usamos os dados do arquivo CRIME1.RAW para verificar se a média de tempo das penas cumpridas de condenações passadas afeta o número de prisões no ano atual (1986). O modelo estimado é

$$\begin{aligned} \widehat{npre86} = {} & 0{,}567 - 0{,}136\,pcond + 0{,}0178\,sentmed - 0{,}00052\,sentmed^2 \\ & \quad (0{,}036)\ (0{,}040) \qquad (0{,}0097) \qquad (0{,}00030) \\ & \quad [0{,}040]\ [0{,}034] \qquad [0{,}0101] \qquad [0{,}00021] \\ & - 0{,}0394\,ptemp86 - 0{,}0505\,empr86 - 0{,}00148\,rend86 \\ & \quad (0{,}0087) \qquad (0{,}0144) \qquad (0{,}00034) \\ & \quad [0{,}0062] \qquad [0{,}0142] \qquad [0{,}00023] \\ & + 0{,}325\,negro + 0{,}193\,hispan \\ & \quad (0{,}045) \qquad (0{,}040) \\ & \quad [0{,}058] \qquad [0{,}040] \\ & n = 2{,}725,\ R^2 = 0{,}0728. \end{aligned} \tag{8.9}$$

Neste exemplo existem diferenças mais substanciais entre alguns dos erros-padrão usuais e os erros-padrão robustos. Por exemplo, a estatística *t* habitual de $sentmed^2$ é aproximadamente −1,73, enquanto a estatística robusta está em torno de −2,48. Assim, $sentmed^2$ é mais significante usando o erro-padrão robusto.

O efeito de *sentmed* sobre *npre86* é um pouco difícil de reconciliar. Como a relação é quadrática, podemos descobrir onde *sentmed* tem um efeito positivo em *npre86* e onde o efeito se torna negativo. O ponto de reversão é 0,0178[2(0,00052)] ≈ 17,12; observe que esta é uma medida de meses. Literalmente, isso significa que *npre86* é positivamente relacionado com *sentmed* quando *sentmed* é menor que 17 meses; então *sentmed* tem o efeito desencorajador esperado após 17 meses.

Para verificar se a média de tempo das penas cumpridas tem um efeito estatisticamente significante sobre *npre86*, devemos testar as hipóteses conjuntas H_0: $\beta_{sentmed} = 0, \beta_{sentmed^2} = 0$. Utilizando a estatística *LM* usual (veja Seção 5.2), obtemos *LM* = 3,54; em uma distribuição qui-quadrada com dois *gl*, isso produz um *p*-valor = 0,170. Assim, não podemos rejeitar H_0 mesmo no nível de 15%. A estatística *LM* robusta em relação à heteroscedasticidade é *LM* = 4,00 (arredondada para duas casas decimais), com um *p*-valor = 0,135. Isso ainda não é uma evidência muito forte contra H_0; *sentmed* não parece ter um efeito forte sobre *npre86*. [A propósito, quando *sentmed* aparece sozinha em (8.9), isto é, sem o termo quadrático, a estatística *t* usual é 0,658, e a estatística *t* robusta é 0,592.]

8.3 O TESTE DA EXISTÊNCIA DE HETEROSCEDASTICIDADE

Os erros-padrão robustos em relação à heteroscedasticidade oferecem um método simples para calcular estatísticas t que sejam assimptoticamente distribuídas como t, haja ou não a presença de heteroscedasticidade. Também já vimos que as estatísticas F e LM robustas em relação à heteroscedasticidade estão disponíveis. A implementação desses testes não exige o conhecimento prévio da presença ou não de heteroscedasticidade. Não obstante, ainda existem algumas boas razões para fazermos alguns testes simples que possam detectar sua presença. Primeiro, como mencionamos na seção anterior, as estatísticas t habituais têm distribuições t exatas sob as hipóteses do modelo linear clássico. Por essa razão, muitos economistas ainda preferem utilizar os erros-padrão MQO usuais e testar as estatísticas informadas, a menos que haja evidência de heteroscedasticidade. Segundo, se houver a presença de heteroscedasticidade, o estimador MQO não mais será o melhor estimador linear não-viesado. Como veremos na Seção 8.4, é possível obter um estimador melhor que o MQO quando a forma da heteroscedasticidade é conhecida.

Muitos testes de heteroscedasticidade têm sido sugeridos ao longo dos anos. Alguns deles, embora tenham a capacidade de detectar a heteroscedasticidade, não testam diretamente a hipótese de que a variância dos erros não depende das variáveis independentes. Nós nos restringiremos a testes mais modernos, que detectam os tipos de heteroscedasticidade que invalidam as estatísticas MQO habituais. Isso também traz o benefício de colocar todos os testes na mesma estrutura.

Como sempre, iniciamos com o modelo linear

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u, \tag{8.10}$$

onde as hipóteses RLM.1 a RLM.4 são mantidas nesta seção. Particularmente, assumimos que $E(u|x_1, x_2, \ldots, x_k) = 0$, de forma que o MQO seja não-viesado e consistente.

Consideramos como hipótese nula que a hipótese RLM.5 é verdadeira:

$$H_0: \text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2. \tag{8.11}$$

Ou seja, assumimos que a hipótese ideal de homoscedasticidade se mantém, e precisamos que os dados nos informem se isso é adequado ou não. Se não pudermos rejeitar (8.11) em um nível de significância suficientemente pequeno, normalmente concluímos que a heteroscedasticidade não será um problema. Porém, lembre-se de que nunca aceitamos H_0; simplesmente não podemos rejeitá-la.

Como estamos assumindo que u tem uma esperança condicional zero, $\text{Var}(u|\boldsymbol{x}) = E(u^2|\boldsymbol{x})$, e assim a hipótese nula de homoscedasticidade é equivalente a

$$H_0: E(u^2|x_1, x_2, \ldots, x_k) = E(u^2) = \sigma^2.$$

Isso mostra que, para testar a violação da hipótese de homoscedasticidade, queremos verificar se u^2 está relacionado (em valor esperado) a uma ou mais das variáveis explicativas. Se H_0 for falsa, o valor esperado de u^2, dadas as variáveis independentes, pode ser virtualmente qualquer função de x_j. Um método simples é assumir uma função linear:

$$u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + v, \tag{8.12}$$

onde v é um termo erro com média zero, dados x_j. Preste bastante atenção na variável dependente nesta equação: ela é o *quadrado* do erro na equação de regressão original, (8.10). A hipótese nula de homoscedasticidade é

$$H_0: \delta_1 = \delta_2 = ... = \delta_k = 0. \tag{8.13}$$

Sob a hipótese nula, freqüentemente é razoável assumir que o erro em (8.12), v, é independente de x_1, x_2, ..., x_k. Então, sabemos da Seção 5.2 que a estatística F ou a estatística LM, sobre a significância geral das variáveis independentes na explicação de u^2, podem ser usadas para testar (8.13). Ambas as estatísticas terão justificação assimptótica, mesmo que u^2 não possa ser normalmente distribuído. (Por exemplo, se u for normalmente distribuído, então u^2/σ^2 é distribuído como χ^2_1.) Se pudéssemos observar u^2 na amostra, então poderíamos calcular com facilidade essa estatística, computando a regressão MQO de u^2 sobre x_1, x_2, ..., x_k, usando todas as n observações.

Como enfatizamos anteriormente, nunca conheceremos os erros efetivos no modelo populacional, mas temos estimativas deles: o resíduo MQO, $\hat{u}_i$ é uma estimativa do erro u_i para a observação i. Assim, podemos estimar a equação

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + ... + \delta_k x_k + \text{erro} \tag{8.14}$$

e calcular as estatísticas F ou LM da significância conjunta de x_1, ..., x_k. Constata-se que o uso dos resíduos MQO em lugar dos erros não afeta a distribuição de amostra grande das estatísticas F ou LM, embora mostrar isso seja bastante complicado.

Ambas as estatísticas F e LM dependem do R-quadrado da regressão (8.14); chamemos isso de $R^2_{\hat{u}2}$ para distingui-lo do R-quadrado na estimação da equação (8.10). Então, a estatística F será

$$F = \frac{R^2_{\hat{u}2}/k}{(1 - R^2_{\hat{u}2})/(n - k - 1)}, \tag{8.15}$$

onde k é o número de regressores em (8.14); este é o mesmo número de variáveis independentes em (8.10). Calcular (8.15) raramente será necessário, já que a maioria dos programas econométricos calcula automaticamente a estatística F da significância geral de uma regressão. A estatística F tem (aproximadamente) uma distribuição $F_{k,\,n-k-1}$ sob a hipótese nula de homoscedasticidade.

A estatística LM para a heteroscedasticidade é simplesmente o tamanho da amostra multiplicado pelo R-quadrado de (8.14):

$$LM = n \cdot R^2_{\hat{u}2}. \tag{8.16}$$

Sob a hipótese nula, a estatística LM é distribuída assimptoticamente como χ^2_k. Isso também é obtido com muita facilidade após computarmos a regressão (8.14).

A versão LM do teste é geralmente chamada **teste de Breusch-Pagan da heteroscedasticidade (teste BP)**. Breusch e Pagan (1980) sugeriram uma forma diferente do teste o qual assume que os erros são normalmente distribuídos. Koenker (1983) sugeriu a forma da estatística LM em (8.16), que é em geral preferida devido a sua maior aplicabilidade.

Resumimos os passos para verificar a existência de heteroscedasticidade usando o teste BP:

O TESTE DE BREUSCH-PAGAN DA HETEROSCEDASTICIDADE

1. Estime o modelo (8.10) por MQO, como usual. Obtenha os resíduos quadrados MQO, $\hat{u}^2$ (um para cada observação).
2. Compute a regressão (8.14). Guarde o R-quadrado desta regressão, $R^2_{\hat{u}^2}$.
3. Construa a estatística F ou a estatística LM e calcule o p-valor (usando a distribuição $F_{k,n-k-1}$ no caso anterior e a distribuição χ^2_k neste último caso). Se o p-valor for suficientemente pequeno, isto é, abaixo do nível de significância selecionado, então rejeitamos a hipótese nula de homoscedasticidade.

Se o teste BP resultar em um p-valor suficientemente pequeno, alguma medida corretiva deve ser tomada. Uma possibilidade é usar os erros-padrão robustos em relação à heteroscedasticidade e as estatísticas de testes discutidas na seção anterior. Outra possibilidade será discutida na Seção 8.4.

EXEMPLO 8.4

(Heteroscedasticidade nas Equações de Preços de Imóveis)

Utilizamos os dados contidos no arquivo HPRICE1.RAW para verificar a existência de heteroscedasticidade em uma equação simples de preços de imóveis. A equação estimada usando os níveis de todas as variáveis é

$$\begin{gathered} \widehat{preço} = -21{,}77 + 0{,}00207\ tamterr + 0{,}123\ arquad + 13{,}85\ qtdorm \\ \qquad (29{,}48)\ (0{,}00064) \qquad (0{,}013) \qquad (9{,}01) \qquad \textbf{(8.17)} \\ n = 88,\ R^2 = 0{,}672. \end{gathered}$$

Esta equação não nos diz *nada* sobre se o erro no modelo populacional é heteroscedástico. Necessitamos fazer a regressão dos resíduos quadrados MQO sobre as variáveis independentes. O R-quadrado da regressão de $\hat{u}^2$ sobre *tamterr*, *arquad* e *qtdorm* é $R^2_{\hat{u}^2} = 0{,}1601$. Com $n = 88$ e $k = 3$, isso produzirá uma estatística F para a significância das variáveis independentes de $F = [0{,}1601/(1 - 0{,}1601)](84/3) \approx 5{,}34$. O p-valor associado é 0,002, o que é forte evidência contra a hipótese nula. A estatística LM é $88(0{,}1601) \approx 14{,}09$; isso dá um p-valor $\approx 0{,}0028$ (usando a distribuição χ^2_3), produzindo essencialmente a mesma conclusão da estatística F. Isso significa que os erros-padrão usuais informados em (8.17) não são confiáveis.

No Capítulo 6 mencionamos que um dos benefícios de usar a forma funcional logarítmica da variável dependente é que a heteroscedasticidade é muitas vezes reduzida. Neste exemplo, coloquemos *preço*, *tamterr* e *arquad* em forma logarítmica, de forma que as elasticidades do *preço* em relação a *tamterr* e *arquad* sejam constantes. A equação estimada é

$$\begin{gathered} \widehat{\log(preço)} = -1{,}30 + 0{,}168\ \log(tamterr) + 0{,}700\ \log(arquad) + 0{,}037\ qtdorm \\ \qquad (0{,}65)\ (0{,}038) \qquad (0{,}093) \qquad (0{,}028) \qquad \textbf{(8.18)} \\ n = 88,\ R^2 = 0{,}643. \end{gathered}$$

Fazendo a regressão dos resíduos quadrados MQO desta regressão sobre log(*tamterr*), log(*arquad*) e *qtdorm* gera $R^2_{\hat{u}^2} = 0{,}0480$. Assim, $F = 1{,}41$ (p-valor = 0,245) e $LM = 4{,}22$ (p-valor = 0,239). Portanto, não podemos rejeitar a hipótese nula de homoscedasticidade no modelo com a forma funcional logarítmica. A ocorrência de menos heteroscedasticidade com a variável dependente em forma logarítmica tem sido observada em muitas aplicações empíricas.

Considere a equação de salários (7.11), na qual você acredita que a variância condicional de log(*salário*) não depende de *educ*, *exper* ou *perm*. Porém, você está preocupado porque a variância de log(*salário*) pode diferir ao longo dos quatro grupos demográficos de homens casados, mulheres casadas, homens solteiros e mulheres solteiras. Que regressão você faria para verificar a existência de heteroscedasticidade? Quais são os graus de liberdade no teste *F*?

Se suspeitarmos que a heteroscedasticidade depende somente de certas variáveis independentes, podemos, com facilidade, modificar o teste de Breusch-Pagan: simplesmente fazemos a regressão de $\hat{u}^2$ sobre quaisquer variáveis independentes que escolhamos e aplicamos o teste F ou LM apropriado. Lembre-se de que os graus de liberdade apropriados dependem do número de variáveis independentes na regressão com $\hat{u}^2$ como variável dependente; o número de variáveis independentes que aparece na equação (8.10) é irrelevante.

Se os resíduos quadrados forem regredidos somente sobre uma única variável independente, o teste de heteroscedasticidade será a estatística t habitual da variável. Uma estatística t significativa sugere que a heteroscedasticidade é um problema.

O Teste de White para a Heteroscedasticidade

No Capítulo 5 mostramos que tanto os erros-padrão como as estatísticas de testes estimados habitualmente por MQO são assimptoticamente válidos, desde que todas as hipóteses de Gauss-Markov se mantenham. Acontece que a hipótese de homoscedasticidade, $\text{Var}(u_1|x_1, ..., x_k) = \sigma^2$, pode ser substituída pela hipótese mais fraca de que o erro quadrado, u^2, é *não-correlacionado* com todas as variáveis independentes (x_j), com os quadrados das variáveis independentes, (x_j^2), e com todos os produtos cruzados ($x_j x_h$ para $j \neq h$). Essa observação motivou White (1980) a propor um teste para a heteroscedasticidade que adiciona quadrados e produtos cruzados de todas as variáveis independentes à equação (8.14). O teste é explicitamente destinado a testar formas de heteroscedasticidade que invalidem os erros-padrão e as estatísticas de testes habituais, estimados por MQO.

Quando o modelo contém $k = 3$ variáveis independentes, o teste de White baseia-se na estimativa de

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_1^2 + \delta_5 x_2^2 + \delta_6 x_3^2 + \delta_7 x_1 x_2 + \delta_8 x_1 x_3 + \delta_9 x_2 x_3 + \text{erro}. \qquad \textbf{(8.19)}$$

Comparado com o teste de Breusch-Pagan, esta equação tem seis regressores a mais. O **teste de White para a heteroscedasticidade** é a estatística LM para testarmos que todos os δ_j na equação (8.19) sejam zero, exceto o intercepto. Assim, nove restrições estão sendo testadas neste caso. Também podemos usar um teste F desta hipótese; ambos os testes têm justificativa assimptótica.

Com somente três variáveis independentes no modelo original, a equação (8.19) tem nove variáveis independentes. Com seis variáveis independentes no modelo original, a regressão de White envolveria geralmente 27 regressores (a menos que alguns sejam redundantes). Essa abundância de regressores é uma fraqueza na forma pura do teste de White: ele usa muitos graus de liberdade para modelos com um número moderado de variáveis independentes.

É possível obter um teste que seja mais facilmente implementado que o teste de White e menos prejudicial quanto aos graus de liberdade. Para criar esse teste, observe que a diferença entre os testes de White e de Breusch-Pagan é que o primeiro inclui os quadrados e os produtos cruzados das variáveis independentes. Podemos obter o mesmo resultado utilizando menos funções das variáveis independentes. Uma sugestão é usar os valores estimados MQO para verificar a existência de heteroscedasticidade. Lembre-se de que os valores estimados são definidos para cada observação i por

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik}.$$

Os valores estimados são apenas funções lineares das variáveis independentes. Se eles forem elevados ao quadrado, obteremos uma função particular de todos os quadrados e produtos cruzados das variáveis independentes. Isso sugere testar a heteroscedasticidade estimando a equação

$$\hat{u}^2 = \delta_0 + \delta_1 \hat{y} + \delta_2 \hat{y}^2 + \text{erro}, \qquad \textbf{(8.20)}$$

onde $\hat{y}$ representa os valores estimados. É importante não confundir $\hat{y}$ com y nesta equação. Usamos os valores estimados porque eles são funções das variáveis independentes (e dos parâmetros estimados); o uso de y em (8.20) não produz um teste válido para a heteroscedasticidade.

Podemos usar as estatísticas F ou LM para a hipótese nula H_0: $\delta_1 = 0, \delta_2 = 0$ na equação (8.20). Isso resultará em duas restrições ao testar a hipótese nula da homoscedasticidade, a despeito do número de variáveis independentes no modelo original. Preservar os graus de liberdade desta maneira é quase sempre uma boa idéia, e também faz com que o teste seja implementado mais facilmente.

Como $\hat{y}$ é uma estimativa do valor esperado de y, dadas as variáveis x_j, usar (8.20) para testar a heteroscedasticidade é útil nos casos em que há suspeita de que a variância muda com o nível do valor esperado, $E(y|\boldsymbol{x})$. O teste a partir de (8.20) pode ser visto como um caso especial do teste de White, já que pode ser mostrado que a equação (8.20) impõe restrições sobre os parâmetros da equação (8.19).

UM CASO ESPECIAL DO TESTE DE WHITE PARA A HETEROSCEDASTICIDADE:

1. Estime o modelo (8.10) por MQO, da maneira habitual. Obtenha os resíduos $\hat{u}$ e os valores estimados $\hat{y}$ do MQO. Calcule os resíduos quadrados $\hat{u}^2$ e os quadrados dos valores estimados $\hat{y}^2$ do MQO.
2. Compute a regressão da equação (8.20). Guarde o R-quadrado desta regressão, $R^2_{\hat{u}^2}$.
3. Construa as estatísticas F ou LM e calcule o p-valor (usando a distribuição $F_{2,n-3}$ no primeiro caso e a distribuição χ^2_2 no último).

EXEMPLO 8.5

(Forma Especial do Teste de White na Equação do Log dos Preços de Imóveis)

Aplicamos o caso especial do teste de White na equação (8.18) na qual usamos a estatística LM. O importante a lembrar é que a distribuição qui-quadrado sempre tem dois gl. A regressão de $\hat{u}^2$ sobre *lpreço*, $(lpreço)^2$, onde *lpreço* representa os valores estimados em (8.18), produz $R^2_{\hat{u}^2} = 0{,}0392$; assim, $LM = 88(0{,}0392) \approx 3{,}45$, e o p-valor $= 0{,}178$. Isso é evidência de heteroscedasticidade mais forte do que a fornecida pelo teste de Breusch-Pagan, mas ainda não podemos rejeitar a homoscedasticidade, mesmo ao nível de 15%.

Antes de deixarmos esta seção, devemos comentar uma importante limitação. Temos interpretado uma rejeição usando um dos testes de heteroscedasticidade como evidência de heteroscedasticidade. Isso está correto desde que mantenhamos as hipóteses RLM.1 a RLM.4. Mas se RLM.3 for violada — especialmente se a forma funcional de $E(y|\boldsymbol{x})$ estiver mal-especificada — então um teste de heteroscedasticidade pode rejeitar H_0, mesmo se $\text{Var}(y|\boldsymbol{x})$ for constante. Por exemplo, se omitirmos um ou mais termos quadráticos em um modelo de regressão ou usarmos o modelo em nível quando deveríamos usar em log, um teste de heteroscedasticidade pode ser significante. Isso tem levado alguns economistas a verem os testes de heteroscedasticidade como testes generalizados de má especificação. Porém, existem testes melhores e mais diretos para testar a má especificação de formas funcionais, e estudaremos alguns deles na Seção 9.1. É melhor usar, primeiro, testes específicos de formas funcionais, já que a má especificação da forma funcional é mais importante que a heteroscedasticidade. Em seguida, uma vez que estejamos satisfeitos com a forma funcional, podemos fazer o teste para verificar a existência de heteroscedasticidade.

8.4 ESTIMAÇÃO DE MÍNIMOS QUADRADOS PONDERADOS

Se for detectada heteroscedasticidade com o uso de um dos testes da Seção 8.3, sabemos da Seção 8.2 que uma reação possível é usar estatísticas robustas em relação à heteroscedasticidade após a estimação MQO. Antes do desenvolvimento das estatísticas robustas em relação à heteroscedasticidade, a resposta à descoberta de heteroscedasticidade era modelar e estimar sua forma específica. Como veremos, isso leva a um estimador mais eficiente que o MQO, e produz estatísticas t e F que têm distribuições t e F. Embora isso pareça atraente, requer mais trabalho de nossa parte, pois temos de ser muito específicos sobre a natureza de qualquer heteroscedasticidade.

A Heteroscedasticidade É Percebida como uma Constante Multiplicativa

Considere que $\boldsymbol{x}$ representa todas as variáveis explicativas na equação (8.10) e assuma que

$$\text{Var}(u|\boldsymbol{x}) = \sigma^2 h(\boldsymbol{x}), \tag{8.21}$$

onde $h(\boldsymbol{x})$ é alguma função das variáveis explicativas que determina a heteroscedasticidade. Como variâncias devem ser positivas, $h(\boldsymbol{x}) > 0$ para todos os valores possíveis das variáveis independentes. Supomos nesta subseção que a função $h(\boldsymbol{x})$ é conhecida. O parâmetro populacional σ^2 é desconhecido, mas teremos condições de estimá-lo a partir de uma amostra de dados.

Para uma extração aleatória da população, podemos escrever $\sigma_i^2 = \text{Var}(u_i|\boldsymbol{x}_i) = \sigma^2 h(\boldsymbol{x}_i) = \sigma^2 h_i$, onde novamente usamos a notação $\boldsymbol{x}_i$ para representar todas as variáveis independentes para as observações i, enquanto h_i muda a cada observação porque as variáveis independentes mudam ao longo das observações. Por exemplo, considere a função simples de poupança

$$poup_i = \beta_0 + \beta_1 renda_i + u_i \tag{8.22}$$

$$\text{Var}(u_i|renda_i) = \sigma^2 renda_i. \tag{8.23}$$

Aqui, $h(x) = h(renda) = renda$: a variância do erro é proporcional ao nível da renda. Isso significa que, conforme a renda aumenta, a variabilidade da poupança cresce. (Se $\beta_1 > 0$, o valor esperado da poupança também aumenta com a renda.) Como *renda* é sempre positiva, a variância na equação (8.23) garantidamente será sempre positiva. O desvio-padrão de u_i, condicional em $renda_i$, é $\sigma\sqrt{renda_i}$.

Como podemos usar a informação da equação (8.21) para estimar β_j? Essencialmente, levamos em conta a equação original,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + u_i, \qquad \textbf{(8.24)}$$

que contém erros heteroscedásticos, e a transformamos em uma equação que não contenha tais erros (e satisfaça as outras hipóteses de Gauss-Markov). Como h_i é apenas uma função de x_i, $u_i/\sqrt{h_i}$ tem zero como valor esperado condicional em $\boldsymbol{x}_i$. Além disso, como $\text{Var}(u_i|\boldsymbol{x}_i) = \text{E}(u_i^2|\boldsymbol{x}_i) = \sigma^2 h_i$, a variância de $u_i/\sqrt{h_i}$ (condicional em $\boldsymbol{x}_i$) é σ^2:

$$\text{E}\left((u_i/\sqrt{h_i})^2\right) = \text{E}(u_i^2)/h_i = (\sigma^2 h_i)/h_i = \sigma^2,$$

onde suprimimos a condicionalidade em $\boldsymbol{x}_i$, por simplicidade. Podemos dividir a equação (8.24) por $\sqrt{h_i}$ para obter

$$\begin{aligned} y_i/\sqrt{h_i} = \beta_0/\sqrt{h_i} + \beta_1(x_{i1}/\sqrt{h_i}) + \beta_2(x_{i2}/\sqrt{h_i}) + \ldots \\ + \beta_k(x_{ik}/\sqrt{h_i}) + (u_i/\sqrt{h_i}) \end{aligned} \qquad \textbf{(8.25)}$$

ou

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \ldots + \beta_k x_{ik}^* + u_i^*, \qquad \textbf{(8.26)}$$

onde $x_{i0}^* = 1/\sqrt{h_i}$ e as outras variáveis sobrescritas com * representam as variáveis originais correspondentes divididas por $\sqrt{h_i}$.

A equação (8.26) parece um pouco peculiar, mas o importante a ser lembrado é que a derivamos para podermos obter os estimadores de β_j que tenham propriedades de eficiência melhores que MQO. O intercepto β_0 na equação original (8.24) agora está multiplicando a variável $x_{i0}^* = 1/\sqrt{h_i}$. Cada parâmetro de inclinação em β_j multiplica uma nova variável que raramente tem interpretação útil. Isso não deve causar problemas se lembrarmos que, na interpretação dos parâmetros e do modelo, sempre queremos retornar à equação original (8.24).

No exemplo precedente da poupança, a equação transformada se assemelha a

$$poup_i/\sqrt{renda_i} = \beta_0(1/\sqrt{renda_i}) + \beta_1\sqrt{renda_i} + u_i^*,$$

onde usamos o fato de que $renda_i/\sqrt{renda_i} = \sqrt{renda_i}$. Contudo, β_1 é a propensão marginal a poupar, uma interpretação que obtemos da equação (8.22).

A equação (8.26) é linear em seus parâmetros (portanto satisfaz RLM.1), e a hipótese de amostragem aleatória não se alterou. Além disso, u_i^* tem uma média zero e uma variância constante (σ^2), condicional em x_i^*. Isso significa que se a equação original satisfizer as quatro primeiras hipóteses de Gauss-Markov, então a equação transformada (8.26) satisfará todas as cinco hipóteses de Gauss-Markov. Além disso, se u_i tiver uma distribuição normal, então u_i^* terá uma distribuição normal com variância σ^2. Portanto, a equação transformada satisfará as hipóteses do modelo linear clássico (RLM.1 a RLM.6), se o modelo original também o fizer, com exceção da hipótese de homoscedasticidade.

Como sabemos que o MQO tem propriedades atraentes (BLUE, por exemplo) sob as hipóteses de Gauss-Markov, a discussão no parágrafo anterior sugere estimarmos os parâmetros da equação (8.26) por mínimos quadrados ordinários. Esses estimadores, $\beta_0^*, \beta_1^*, \ldots, \beta_k^*$, serão diferentes dos estimadores MQO na equação original. Os β_j^* são exemplos de **estimadores de mínimos quadrados generalizados (MQG)**. Neste caso, os estimadores MQG são usados para explicar a heteroscedasticidade nos erros. Encontraremos outros estimadores MQG no Capítulo 12.

Como a equação (8.26) satisfaz todas as hipóteses ideais, erros-padrão, estatísticas t e estatísticas F podem ser obtidas de regressões que usem as variáveis transformadas. A soma dos quadrados dos resíduos em (8.26) dividida pelos graus de liberdade é um estimador não-viesado de σ^2. Além disso, os estimadores MQG, por serem os melhores estimadores lineares não-viesados de β_j, são necessariamente mais eficientes que os estimadores MQO $\hat{\beta}_j$ obtidos da equação não transformada. Em essência, após termos transformado as variáveis, simplesmente usamos a análise padrão MQO. Porém, devemos nos lembrar de interpretar as estimativas à luz da equação original.

O R-quadrado obtido da estimação em (8.26), embora seja útil para computar estatísticas F, não é especialmente informativo como uma medida do grau de ajuste: ele nos informa quanto da variação em y^* é explicado por x_j^*, e isso raramente é significativo.

Os estimadores MQG para a correção da heteroscedasticidade são chamados de **estimadores de mínimos quadrados ponderados (MQP)**. Esse nome advém do fato de que β_j^* minimiza a soma *ponderada* dos quadrados dos resíduos, onde cada resíduo quadrado é ponderado por $1/h_i$. A idéia é colocar menos peso nas observações com uma variância de erro mais alta; o método MQO dá a cada observação o mesmo peso, pois isso é melhor quando a variância do erro é idêntica para todas as partições da população. Matematicamente, os estimadores MQP são os valores de b_j que tornam a expressão

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{ik})^2/h_i \qquad \textbf{(8.27)}$$

tão pequena quanto possível. Levar a raiz quadrada de $1/h_i$ para dentro do resíduo quadrado mostra que a soma ponderada dos quadrados dos resíduos é idêntica à soma dos quadrados dos resíduos nas variáveis transformadas:

$$\sum_{i=1}^{n} (y_i^* - b_0 x_{i0}^* - b_1 x_{i1}^* - b_2 x_{i2}^* - \ldots - b_k x_{ik}^*)^2.$$

Como MQO minimiza a soma dos resíduos quadrados (a despeito das definições das variáveis dependente e independentes), concluímos que os estimadores MQP que minimizam (8.27) são simplesmente os estimadores MQO de (8.26). Observe cuidadosamente que os resíduos quadrados em (8.27) são ponderados por $1/h_i$, enquanto as variáveis transformadas em (8.26) são ponderadas por $1/\sqrt{h_i}$.

Um estimador de mínimos quadrados ponderados pode ser definido para qualquer conjunto de pesos positivos. O MQO é o caso especial que atribui pesos iguais a todas as observações. O procedi-

mento eficiente, MQG, pondera cada resíduo quadrado pelo inverso da variância condicional de u_i, dado x_i.

A obtenção das variáveis transformadas na equação (8.25) com o objetivo de calcular manualmente os mínimos quadrados ponderados pode ser entediante, e não se pode desprezar a possibilidade de cometer erros. Felizmente, a maioria dos programas econométricos modernos tem um recurso para computar mínimos quadrados ponderados. Em geral, juntamente com as variáveis dependentes e independentes no modelo original, apenas especificamos a função de ponderação, $1/h_i$, que aparece em (8.27). Isto é, especificamos pesos proporcionais ao inverso da variância, e não ao desvio-padrão. Além de haver menor possibilidade de cometermos erros, isso nos força a interpretar as estimativas de mínimos quadrados ponderados no modelo original. Aliás, podemos escrever a equação estimada da maneira habitual. As estimativas e os erros-padrão serão diferentes do MQO, mas a maneira como interpretamos essas estimativas, erros-padrão e estatísticas de testes é a mesma.

EXEMPLO 8.6

(Equação de Poupança Familiar)

A Tabela 8.1 contém estimativas de funções de poupança construídas a partir do arquivo de dados SAVING.RAW (sobre 100 famílias, desde 1970). Estimamos o modelo de regressão simples (8.22) por MQO e por mínimos quadrados ponderados, assumindo no último caso que a variância é dada por (8.23). Em seguida, adicionamos as variáveis tamanho da família, idade do chefe da família, anos de escolaridade do chefe da família e uma variável *dummy* indicando se o chefe da família é negro.

No modelo de regressão simples, a estimativa MQO da propensão marginal a poupar (PMP) é 0,147, com uma estatística t de 2,53. (Os erros-padrão MQO na Tabela 8.1 são os erros-padrão não-robustos. Se realmente considerássemos que a heteroscedasticidade seria um problema, provavelmente computaríamos também os erros-padrão robustos em relação à heteroscedasticidade, o que não faremos neste caso.) A estimativa MQP da PMP é um pouco mais alta: 0,172, com $t = 3,02$. Os erros-padrão das estimativas MQO e MQP são muito semelhantes para este coeficiente. As estimativas MQO e MQP do intercepto são bastante diferentes, mas isso não deve causar preocupação já que as estatísticas t são ambas muito pequenas. Encontrar mudanças razoavelmente grandes em coeficientes que não são significantes não é incomum quando se comparam as estimativas MQO e MQP. Os R-quadrados nas colunas (1) e (2) não são comparáveis.

A adição de variáveis demográficas reduz a PMP, quer usemos o MQO ou o MQP; os erros-padrão também aumentam em quantidade razoável (devido à multicolinearidade que é induzida pela inclusão dessas variáveis adicionais). É fácil verificar, usando as estimativas do MQO ou do MQP, que nenhuma das variáveis adicionais é, individualmente, significativa. Serão elas significativas conjuntamente? O teste F baseado na estimativa MQO usa os R-quadrados das colunas (1) e (3). Com 94 *gl* no modelo irrestrito e com quatro restrições, a estatística F é $F = [(0,0828 - 0,0621)/(1 - 0,0828)](94/4) \approx 0,53$ e o p-valor $= 0,715$. O teste F, usando as estimativas MQP, usa os R-quadrados das colunas (2) e (4): $F \approx 0,50$ e o p-valor $= 0,739$. Assim, usando o MQO ou o MQP, as variáveis demográficas são conjuntamente não significantes. Isso sugere que o modelo de regressão simples relacionando poupança à renda é suficiente.

Qual deveríamos escolher como nossa melhor estimativa da propensão marginal a poupar? Neste caso, não importa muito se usarmos a estimativa MQO de 0,147 ou a estimativa MQP de 0,172. Lembre-se de que ambas são apenas estimativas de uma amostra relativamente pequena, e o intervalo de confiança de 95% do MQO contém a estimativa MQP, e vice-versa.

Tabela 8.1

Variável Dependente: poup

Variáveis Independentes	(1) MQO	(2) MQP	(3) MQO	(4) MQP
renda	0,147 (0,058)	0,172 (0,057)	0,109 (0,071)	0,101 (0,077)
tamanho	——	——	67,66 (222,96)	−6,87 (168,43)
educ	——	——	151,82 (117,25)	139,48 (100,54)
idade	——	——	0,286 (50,031)	21,75 (41,31)
negro	——	——	518,39 (1.308,06)	137,28 (844,59)
intercepto	124,84 (655,39)	−124,95 (480,86)	−1.605,42 (2.830,71)	−1.854,81 (2.351,80)
Observações *R-quadrado*	100 0,0621	100 0,0853	100 0,0828	100 0,1042

Usando os resíduos MQO obtidos da regressão MQO descrita na coluna (1) da Tabela 8.1, a regressão de $\hat{u}^2$ sobre *renda* produz uma estatística *t* do coeficiente de *renda* de 0,96. Existe alguma necessidade de se usar mínimos quadrados ponderados no Exemplo 8.6?

Na prática, raramente sabemos na forma simples como a variância depende de uma determinada variável independente. Por exemplo, na equação de poupança que inclui todas as variáveis demográficas, como sabemos que a variância de *poup* não muda com a idade ou os níveis de educação? Na maioria das aplicações ficamos inseguros sobre $\text{Var}(y|x_1, x_2, ..., x_k)$.

Existe um caso no qual os pesos necessários para o MQP surgem naturalmente de um modelo econométrico subjacente. Isso acontece quando, em vez de usarmos dados em nível individual, somente temos médias de dados de algum grupo ou região geográfica. Por exemplo, suponha que estejamos interessados em determinar a relação entre o montante que um trabalhador contribui para seu plano de pensão como uma função da generosidade do plano. Suponhamos que *i* seja uma empresa em particular e que *e* represente um empregado dessa empresa. Um modelo simples é

$$contrib_{i,e} = \beta_0 + \beta_1 ganhos_{i,e} + \beta_2 idade_{i,e} + \beta_3 taxcont_i + u_{i,e}, \qquad \textbf{(8.28)}$$

onde $contrib_{i,e}$ é a contribuição anual por empregado e que trabalha na empresa i, $ganhos_{i,e}$ é o ganho anual dessa pessoa, e $idade_{i,e}$ é a idade dessa pessoa. A variável $taxcont_i$ é o montante que a empresa deposita na conta do empregado para cada dólar pago em contribuição pelo empregado.

Se (8.28) satisfizer as hipóteses de Gauss-Markov, então poderemos estimá-la, dada uma amostra de indivíduos entre vários empregadores. Suponha, porém, que somente temos valores *médios* de contribuições, ganhos e idade, por empregador. Em outras palavras, dados em nível individual não estão disponíveis. Assim, façamos com que $\overline{contrib}_i$ represente a contribuição média do pessoal da empresa i e, semelhantemente, $\overline{ganhos}_i$ e $\overline{idade}_i$ também representem médias. Seja m_i o número de empregados da empresa i; assumimos que esse é um número conhecido. Então, se computarmos a média da equação (8.28) para todos os empregados da empresa i, obteremos a equação dela

$$\overline{contrib}_i = \beta_0 + \beta_1 \overline{ganhos}_i + \beta_2 \overline{idade}_i + \beta_3 taxcont_i + \bar{u}_i, \qquad \textbf{(8.29)}$$

onde $\bar{u}_i = m_i^{-1} \sum_{e=1}^{m_i} u_{i,e}$ é o erro médio entre todos os empregados da empresa i. Se tivermos n empresas na nossa amostra, então (8.29) será apenas um modelo de regressão linear múltipla que pode ser estimado por MQO. Os estimadores serão não-viesados se o modelo original (8.28) satisfizer as hipóteses de Gauss-Markov e os erros individuais $u_{i,e}$ serão independentes do tamanho da empresa, m_i [porque então o valor esperado de $\bar{u}_i$, dadas as variáveis explicativas em (8.29), será zero].

Se a equação no nível individual satisfizer a hipótese de homoscedasticidade, então a equação da empresa (8.29) deverá ter heteroscedasticidade. Assim, se $\text{Var}(u_{i,e}) = \sigma^2$ para todo i e e, então $\text{Var}(\bar{u}_i) = \sigma^2/m_i$. Em outras palavras, para empresas maiores, a variância do termo erro $\bar{u}_i$ diminui com o tamanho da empresa. Neste caso, $h_i = 1/m_i$, e, portanto, o procedimento mais eficiente será o dos mínimos quadrados ponderados, com pesos correspondentes ao número de empregados das empresas ($1/h_i = m_i$). Isso garante que empresas grandes recebam maior peso, o que nos oferece um método eficiente de estimação dos parâmetros no modelo em nível individual quando somente temos médias para as empresas.

Uma ponderação semelhante surge quando estamos usando dados *per capita* no nível de cidade, município, estado ou país. Se a equação no nível individual satisfizer as hipóteses de Gauss-Markov, o erro na equação *per capita* terá uma variância proporcional a um sobre o tamanho da população. Portanto, mínimos quadrados ponderados com pesos iguais à população serão apropriados. Por exemplo, suponha que temos dados em nível de cidades sobre o consumo *per capita* de cerveja (em onças*), a percentagem de pessoas com mais de 21 anos na população, níveis médios de educação dos adultos, níveis médios de renda e o preço da cerveja nas cidades. Então, o modelo no nível de cidades

$$cervpc = \beta_0 + \beta_1 perc21 + \beta_2 educmed + \beta_3 rendapc + \beta_4 preço + u$$

pode ser estimado por mínimos quadrados ponderados, com pesos iguais à população das cidades.

A vantagem de fazer a ponderação pelos tamanhos da empresa, população da cidade, e assim por diante, depende de a equação individual subjacente ser homoscedástica. Se existir heteroscedasticidade no nível individual, então a ponderação adequada dependerá da forma da heteroscedasticidade. Esta é uma razão que explica porque um número cada vez maior de pesquisadores simplesmente computa erros-padrão e estatísticas de teste robustos na estimação de modelos que usam dados *per capita*. Uma

* NT: 1 onça = 29,574 mililitros.

alternativa é ponderar pela população, mas registrar as estatísticas robustas em relação à heteroscedasticidade na estimação MQP. Isso assegura que, embora a estimação seja eficiente se o modelo no nível individual satisfizer as hipóteses de Gauss-Markov, qualquer heteroscedasticidade no nível individual seja representada pela inferência robusta.

A Necessidade de Estimar a Função de Heteroscedasticidade: O MQG Factível

Na subseção anterior, vimos alguns exemplos nos quais a heteroscedasticidade é conhecida como uma forma multiplicativa. Na maioria dos casos, a forma exata de heteroscedasticidade não é óbvia. Em outras palavras, é difícil encontrar a função $h(\boldsymbol{x}_i)$ da seção anterior. Contudo, em muitos casos podemos modelar a função h e utilizar os dados para estimar os parâmetros desconhecidos nesse modelo. Isso resulta em uma estimativa de cada h_i, indicada por $\hat{h}_i$. O uso de $\hat{h}_i$ em lugar de h_i na transformação MQG produz um estimador chamado **estimador MQG factível (MQGF)**. O MQG factível algumas vezes é chamado de *MQG estimado* ou MQGE.

Existem várias maneiras de modelar a heteroscedasticidade, mas estudaremos um método particular razoavelmente flexível. Assuma que

$$\text{Var}(u|\boldsymbol{x}) = \sigma^2\exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k), \qquad \textbf{(8.30)}$$

onde x_1, x_2, ..., x_k são variáveis independentes que aparecem no modelo de regressão [veja equação (8.1)], e δ_j são parâmetros desconhecidos. Outras funções de x_j podem aparecer, mas nos concentraremos primariamente em (8.30). Na notação da subseção anterior, $h(\boldsymbol{x}) = \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k)$.

Você pode estar curioso para saber por que usamos a função exponencial em (8.30). Afinal de contas, quando fizemos o teste para verificar a existência de heteroscedasticidade usando o teste de Breusch-Pagan, assumimos que a heteroscedasticidade era uma função linear de x_j. Alternativas lineares como (8.12) são boas quando verificamos a existência de heteroscedasticidade, mas elas podem ser problemáticas quando fazemos a correção da heteroscedasticidade usando mínimos quadrados ponderados. Já encontramos a razão para esse problema antes: modelos lineares não asseguram que os valores previstos sejam positivos, e nossas variâncias estimadas devem ser positivas para podermos usar o método MQP.

Se os parâmetros δ_j fossem conhecidos, apenas aplicaríamos o MQP, como na subseção anterior. Isso não é muito realista. É melhor usar os dados para estimar esses parâmetros, e então utilizar essas estimativas para construir os pesos. Como podemos estimar os δ_j? Basicamente, transformaremos essa equação em uma forma linear que, com pequenas modificações, poderá ser estimada por MQO.

Sob a hipótese (8.30), podemos escrever

$$u^2 = \sigma^2\exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k)v,$$

onde v tem uma média igual à unidade, condicional em $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$. Se assumirmos que v é realmente independente de $\boldsymbol{x}$, podemos escrever

$$\log(u^2) = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \ldots + \delta_k x_k + e, \qquad \textbf{(8.31)}$$

onde e tem média zero e é independente de x; o intercepto nessa equação é diferente de δ_0, mas isso não é importante. A variável dependente é o log do erro quadrado.

Como (8.31) satisfaz as hipóteses de Gauss-Markov, podemos obter estimadores não-viesados de δ_j utilizando MQO.

Como sempre, devemos substituir o u não-observado pelos resíduos MQO. Portanto, computamos a regressão de

$$\log(\hat{u}^2) \text{ sobre } x_1, x_2, ..., x_k. \qquad \textbf{(8.32)}$$

Na realidade, o que necessitamos dessa regressão são os valores estimados; vamos chamá-los de $\hat{g}_i$. Então, as estimativas de h_i serão simplesmente

$$\hat{h}_i = \exp(\hat{g}_i). \qquad \textbf{(8.33)}$$

Agora usamos o método MQP com pesos $1/\hat{h}_i$ em lugar de $1/h_i$ na equação (8.27). Façamos um resumo dos passos.

UM PROCEDIMENTO MQG FACTÍVEL PARA CORRIGIR A HETEROSCEDASTICIDADE

1. Execute a regressão de y sobre $x_1, x_2, ..., x_k$ e obtenha os resíduos $\hat{u}$.
2. Crie $\log(\hat{u}^2)$ primeiramente elevando ao quadrado os resíduos MQO e depois calculando seu log natural.
3. Execute a regressão na equação (8.32) e obtenha os valores estimados, $\hat{g}$.
4. Calcule o exponencial dos valores estimados a partir de (8.32): $\hat{h} = \exp(\hat{g})$.
5. Estime a equação

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$$

pelo método MQP, usando pesos $1/\hat{h}$.

Se pudéssemos usar h_i em vez de $\hat{h}_i$ no procedimento MQP, saberíamos que nossos estimadores seriam não-viesados; de fato, eles seriam os melhores estimadores lineares não-viesados, supondo que tenhamos modelado apropriadamente a heteroscedasticidade. Estimar h_i usando os mesmos dados significa que o estimador MQGV deixa de ser não-viesado (portanto, ele tampouco pode ser BLUE). No entanto, o estimador MQGV é consistente e assimptoticamente mais eficiente que o MQO. Isso é difícil de demonstrar devido à estimação dos parâmetros da variância. Entretanto, se ignorarmos isso — o que, no fim das contas, faremos — a prova é semelhante à demonstração de que o MQO é eficiente na classe de estimadores no Teorema 5.3. De qualquer forma, para amostras de tamanho grande, o MQGV é uma atraente alternativa ao MQO quando existe evidência de heteroscedasticidade que infla os erros-padrão das estimativas MQO.

Devemos lembrar que os estimadores MQGV são estimadores dos parâmetros na equação

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u.$$

Assim como as estimativas MQO medem o impacto marginal de cada x_j sobre y, as estimativas MQGV também têm essa característica. Usamos as estimativas MQGV em lugar das estimativas MQO porque

elas são mais eficientes e possuem estatísticas de testes associadas às distribuições t e F usuais, pelo menos em amostras grandes. Se tivermos alguma dúvida sobre a variância especificada na equação (8.30), poderemos usar os erros-padrão e as estatísticas de testes robustos em relação à heteroscedasticidade na equação transformada.

Outra alternativa útil para estimar h_i é substituir as variáveis independentes na regressão (8.32) pelos valores estimados por MQO e seus quadrados. Em outras palavras, obter os $\hat{g}_i$ como os valores estimados da regressão de

$$\log(\hat{u}^2) \text{ sobre } \hat{y}, \hat{y}^2 \qquad \textbf{(8.34)}$$

e depois obter os $\hat{h}_i$ exatamente como na equação (8.33). Isso altera apenas o passo (3) do procedimento anterior.

Se usarmos a regressão (8.32) para estimar a função da variância, você vai querer saber se podemos simplesmente fazer o teste para verificar a existência de heteroscedasticidade usando essa mesma regressão (um teste F ou LM pode ser usado). Aliás, Park (1966) fez essa sugestão. Infelizmente, quando comparado com os testes discutidos na Seção 8.3, o teste de Park tem alguns problemas. Primeiro, a hipótese nula deve ser algo mais forte que a homoscedasticidade: efetivamente, u e x devem ser independentes. Isso não é exigido pelos testes de Breusch-Pagan ou White. Segundo, o uso dos resíduos MQO $\hat{u}$ em lugar de u em (8.32) pode fazer com que a estatística F se desvie da distribuição F, mesmo em amostras de tamanho grande. Isso não é um problema nos outros testes que tratamos. Por essas razões, o teste de Park não é recomendado quando estivermos fazendo testes para verificar a existência de heteroscedasticidade. A razão pela qual a regressão (8.32) funciona bem para mínimos quadrados ponderados é que somente precisamos de estimadores consistentes de δ_j, e a regressão (8.32) certamente os produz.

EXEMPLO 8.7

(Demanda de Cigarros)

Utilizamos os dados contidos no arquivo SMOKE.RAW para estimar uma função de demanda de consumo diário de cigarros. Como a maioria das pessoas não fuma, a variável dependente *cigs* é zero para a maioria das observações. Um modelo linear não é o ideal, pois ele pode produzir valores previstos negativos. Mesmo assim, podemos aprender alguma coisa sobre os determinantes do hábito de fumar utilizando um modelo linear.

A equação estimada por mínimos quadrados ponderados, com os erros-padrão MQO usuais entre parênteses, é

$$\begin{aligned} \widehat{cigs} = &-3{,}64 + 0{,}880\log(renda) - 0{,}751\log(precig) \\ &\quad (24{,}08)\ (0{,}728)\qquad\quad (5{,}773) \\ &-0{,}501\,educ + 0{,}771\,idade - 0{,}0090\,idade^2 - 2{,}83\,restaurn \\ &\ (0{,}167)\qquad (0{,}160)\qquad\ (0{,}0017)\qquad\ (1{,}11) \\ &\qquad n = 807,\ R^2 = 0{,}0526, \end{aligned} \qquad \textbf{(8.35)}$$

EXEMPLO 8.7 (continuação)

onde *cigs* é o número de cigarros fumados por dia, *renda* é a renda anual, *precig* é o preço por maço de cigarros (em centavos de dólar), *educ* representa anos de escolaridade formal, *idade* é medida em anos e *restaurn* é um indicador binário igual a um se a pessoa residir em um estado com restrições de fumar em restaurantes. Como também vamos trabalhar com o método de mínimos quadrados ponderados, não registraremos os erros-padrão robustos em relação à heteroscedasticidade do MQO. (A propósito, 13 dos 807 valores estimados são menores que zero; isso é menos de 2% da amostra e não é motivo importante para preocupação.)

Nem a renda nem o preço dos cigarros são estatisticamente significantes em (8.35), e seus efeitos não são grandes na prática. Por exemplo, se a renda aumenta em 10%, *cigs* aumenta previsivelmente em (0,880/100)(10) = 0,088, ou menos de um décimo de cigarro por dia. A magnitude do efeito do preço é semelhante.

Cada ano de educação formal reduz a média de cigarros fumados por dia à metade, e o efeito é estatisticamente significante. O hábito de fumar também está relacionado com a idade, de um modo quadrático, aumenta com a idade até a *idade* = 0,771/[2(0,009)] ≈ 42,83, e depois diminui com a idade. Os dois termos relacionados à idade são estatisticamente significantes. A presença de uma restrição quanto a fumar em restaurantes reduz o hábito de fumar em quase três cigarros por dia, em média.

Os erros subjacentes na equação (8.35) contêm heteroscedasticidade? A regressão de Breusch-Pagan dos resíduos quadrados MQO sobre as variáveis independentes em (8.35) [veja equação (8.14)] produz $R^2_{\hat{u}^2} = 0,040$. Esse R-quadrado pequeno pode parecer indicar a não-existência de heteroscedasticidade, mas devemos nos lembrar de calcular a estatística F ou a LM. Se o tamanho da amostra for grande, um $R^2_{\hat{u}^2}$ aparentemente pequeno pode resultar em uma rejeição muito forte da homoscedasticidade. A estatística LM é 807(0,040) = 32,28, e esse é o resultado de uma variável aleatória χ^2_6. O p-valor é menor que 0,000015, o que é uma evidência muito forte de heteroscedasticidade.

Portanto, estimamos a equação usando o procedimento MQG factível anterior. A equação estimada é

$$\begin{gathered}\widehat{cigs} = 5,64 + 1,30\log(renda) - 2,94\log(precig) \\ (17,80)\ \ (0,44)\qquad\qquad (4,46) \\ -0,463\,educ + 0,482\,idade - 0,0056\,idade^2 - 3,46\,restaurn \\ (0,120)\qquad (0,097)\qquad (0,0009)\qquad (0,80) \\ n = 807,\ R^2 = 0,1134.\end{gathered} \tag{8.36}$$

O efeito da renda agora é estatisticamente significante e maior em magnitude. O efeito do preço é, também, notavelmente maior, mas continua sendo estatisticamente não significante. [Uma razão para isso é que *precig* varia somente entre estados na amostra, e assim existe muito menos variação em log(*precig*) do que em log(*renda*), *educ* e *idade*.]

As estimativas sobre as outras variáveis, naturalmente, mudaram um pouco, mas a história continua a mesma. Fumar está relacionado negativamente com escolaridade, tem uma relação quadrática com a idade, e é negativamente afetado pelas restrições de fumar em restaurantes.

Devemos ter algum cuidado ao calcularmos estatísticas F para testar hipóteses múltiplas após a estimação por MQP. (Isso é válido se a soma dos resíduos quadrados ou a forma R-quadrada da estatística F for usada.) É importante que os mesmos pesos sejam usados para estimar os modelos com e

sem restrições. Devemos primeiro estimar o modelo sem restrições por MQO. Uma vez que tenhamos obtido os pesos, poderemos usá-los para também estimar o modelo restrito. A estatística F pode ser calculada da maneira habitual. Felizmente, muitos programas econométricos possuem um comando simples para testar restrições conjuntas após a estimação MQP, de modo que não precisamos, nós mesmos, calcular manualmente a regressão restrita.

Suponha que o modelo da heteroscedasticidade na equação (8.30) não esteja correto, mas usamos o procedimento MQG factível baseado nessa variância. O MQP ainda é coerente, mas os usuais erros-padrão, estatísticas t etc. não serão válidos, mesmo assimptoticamente. O que podemos fazer? [*Sugestão*: Veja a equação (8.26), na qual u_i^* contém heteroscedasticidade se $\text{Var}(u|\boldsymbol{x}) \neq \sigma^2 h(\boldsymbol{x})$.]

O Exemplo 8.7 sugere uma questão que às vezes surge em aplicações de mínimos quadrados ponderados: as estimativas MQO e MQP podem ser substancialmente diferentes. Isso não é um grande problema na equação de demanda de cigarros, pois todos os coeficientes mantêm os mesmos sinais, e as maiores mudanças são nas variáveis que eram estatisticamente não significantes quando a equação foi estimada por MQO. As estimativas MQO e MQP sempre serão diferentes devido ao erro amostral. O problema é quando suas diferenças são suficientes para alterar conclusões importantes.

Se os métodos MQO e MQP produzirem estimativas estatisticamente significantes que sejam diferentes nos sinais — por exemplo, a elasticidade-preço estimada por MQO é significante e positiva, enquanto a estimada por MQP é significante e negativa — ou se as diferenças em magnitude das estimativas forem de fato grandes, devemos ficar desconfiados. Em geral, isso indica que uma das outras hipóteses de Gauss-Markov é falsa, particularmente a hipótese de média condicional dos erros nula (RLM.3). A correlação entre u e qualquer variável independente causa viés e inconsistência no MQO e no MQP, e os vieses normalmente serão diferentes. O *teste de Hausman* [Hausman (1978)] pode ser usado para comparar formalmente as estimativas MQO e MQP para verificar se elas diferem mais do que é sugerido pelo erro amostral. Esse teste está além do escopo deste livro. Em muitos casos, um exame informal das estimativas é suficiente para detectar o problema.

8.5 O MODELO DE PROBABILIDADE LINEAR REVISITADO

Como vimos na Seção 7.5, quando a variável dependente y é binária, o modelo deve conter heteroscedasticidade, a menos que todos os parâmetros de inclinação sejam nulos. Estamos agora em posição de lidar com esse problema.

A maneira mais simples de tratar a heteroscedasticidade no modelo de probabilidade linear é continuar a usar a estimação MQO, mas também calcular os erros-padrão robustos nas estatísticas de testes. Isso ignora o fato de que efetivamente conhecemos a forma da heteroscedasticidade do MPL. Contudo, as estimativas MQO do MPL são simples e geralmente produzem resultados satisfatórios.

EXEMPLO 8.8

(Participação de Mulheres Casadas na Força de Trabalho)

No exemplo da participação na força de trabalho na Seção 7.5 [veja equação (7.29)], registramos os erros-padrão MQO habituais. Agora, também computamos os erros-padrão robustos em relação à heteroscedasticidade. Eles estão registrados entre colchetes abaixo dos erros-padrão usuais:

$$\begin{aligned} \widehat{naft} = {} & 0{,}586 - 0{,}0034\ nesprend + 0{,}038\ educ + 0{,}039\ exper \\ & \quad (0{,}154)\ \ (0{,}0014)\ \ \ (0{,}007)\ \ \ (0{,}006) \\ & \quad [0{,}151]\ [0{,}0015]\ \ [0{,}007]\ \ [0{,}006] \\ & - 0{,}00060\ exper^2 - 0{,}016\ idade - 0{,}262\ crianme6 + 0{,}0130\ crianma6 \\ & \quad (0{,}00018)\ \ \ (0{,}002)\ \ \ (0{,}034)\ \ \ (0{,}0132) \\ & \quad [0{,}00019]\ \ \ [0{,}002]\ \ \ [0{,}032]\ \ \ [0{,}0135] \\ & \qquad n = 753,\ R^2 = 0{,}264. \end{aligned} \tag{8.37}$$

Vários dos erros-padrão robustos e MQO são os mesmos para o grau de precisão registrado; em todos os casos, as diferenças são de fato muito pequenas. Portanto, embora a heteroscedasticidade seja um problema na teoria, ela não é na prática, pelo menos neste exemplo. Muitas vezes constata-se que os erros-padrão e estatísticas de testes usuais MQO são semelhantes aos seus correspondentes robustos em relação à heteroscedasticidade. E mais ainda, o esforço exigido para computar ambos é mínimo.

Geralmente, os estimadores MQO são ineficientes no MPL. Lembre-se de que a variância condicional de y no MPL é

$$\text{Var}(y|\boldsymbol{x}) = p(\boldsymbol{x})[1 - p(\boldsymbol{x})], \tag{8.38}$$

onde

$$p(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \tag{8.39}$$

é a probabilidade de resposta (probabilidade de sucesso, $y = 1$). Parece natural usar mínimos quadrados ponderados, mas existem alguns obstáculos. A probabilidade $p(\boldsymbol{x})$ claramente depende de parâmetros desconhecidos da população, β_j. No entanto, realmente temos estimadores não-viesados desses parâmetros, ou seja, os estimadores MQO. Quando os estimadores MQO são integrados na equação (8.39), obtemos os valores estimados do MQO. Assim, para cada observação i, $\text{Var}(y_i|\boldsymbol{x}_i)$ é estimada por

$$\hat{h}_i = \hat{y}_i(1 - \hat{y}_i), \tag{8.40}$$

onde $\hat{y}_i$ é o valor estimado por MQO da observação i. Agora, aplicamos o MQG factível, exatamente como na Seção 8.4.

Infelizmente, ter condição de estimar h_i de cada i não significa que poderemos prosseguir diretamente com a estimação MQP. O problema é o que discutimos brevemente na Seção 7.5: os valores estimados $\hat{y}_i$ não precisam cair no intervalo unitário. Se $\hat{y}_i < 0$ ou $\hat{y}_i > 1$, a equação (8.40) mostra que $\hat{h}_i$ será negativo. Como o método MQP prossegue multiplicando a observação i por $1/\sqrt{\hat{h}_i}$, o método falhará se $\hat{h}_i$ for negativo (ou zero) em qualquer observação. Em outras palavras, todos os pesos do método MQP devem ser positivos.

Em alguns casos, $0 < \hat{y}_i < 1$ para todos os i, quando o MQP pode ser usado para estimar o MPL. Nos casos com muitas observações e pequenas probabilidades de sucesso ou fracasso, é muito comum encontrarmos alguns valores estimados fora do intervalo unitário. Se isso acontecer, como no exemplo da participação na força de trabalho na equação (8.37), é mais fácil abandonar o método MQP e registrar as estatísticas robustas em relação à heteroscedasticidade. Uma alternativa é ajustar os valores estimados menores que zero ou maiores que a unidade, e depois aplicar o método MQP. Uma sugestão é definir $\hat{y}_i = 0{,}01$ se $\hat{y}_i < 0$ e $\hat{y}_i = 0{,}99$ se $\hat{y}_i > 1$. Infelizmente, isso exigirá uma escolha arbitrária por parte do pesquisador — por exemplo, por que não usar 0,001 e 0,999 como os valores estimados? Se muitos valores estimados estiverem fora do intervalo unitário, o ajuste pode afetar os resultados; nesta situação, provavelmente será melhor usar somente o método MQO.

ESTIMANDO O MODELO DE PROBABILIDADE LINEAR POR MÍNIMOS QUADRADOS PONDERADOS

1. Estime o modelo por MQO e obtenha os valores estimados $\hat{y}$.
2. Determine se todos os valores estimados estão dentro do intervalo unitário. Se assim for, prossiga para o passo (3). Caso contrário, alguns ajustes serão necessários para trazer todos os valores estimados para dentro do intervalo unitário.
3. Construa as variâncias estimadas na equação (8.40).
4. Estime a equação

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

por MQP, usando pesos $1/\hat{h}$.

EXEMPLO 8.9

(Determinantes da Propriedade de Computadores Pessoais)

Utilizamos os dados contidos no arquivo GPA1.RAW para estimar a probabilidade de possuir um computador. *PC* é um indicador binário igual à unidade se o aluno possuir um computador, e zero caso contrário. A variável *nmem* é a nota média do ensino médio, *tac* é a nota do teste de avaliação de conhecimentos para ingresso em curso superior e *paisfac* é um indicador binário igual a um se pelo menos um dos genitores tem curso superior. (Indicadores separados para a mãe e para o pai não produzem resultados individualmente significantes, já que são bastante correlacionados.)

A equação estimada por MQO é

$$\begin{aligned} \widehat{PC} = {} & -0{,}0004 + 0{,}065\, nmem + 0{,}0006\, ACT + 0{,}221\, paisfac \\ & (0{,}4905)\ (0{,}137)\ \ (0{,}0155)\ \ (0{,}093) \\ & [0{,}4888]\ [0{,}139]\ \ [0{,}0158]\ \ [0{,}087] \\ & n = 141,\ R^2 = 0{,}0415. \end{aligned} \qquad \textbf{(8.41)}$$

EXEMPLO 8.9 (continuação)

Como no exemplo 8.8, não existem diferenças substanciais entre os erros-padrão usuais e os robustos. Não obstante, também estimamos o modelo por MQP. Como todos os valores estimados se encontram dentro do intervalo unitário, não são necessários ajustes:

$$\widehat{PC} = 0{,}026 + 0{,}033\, nmem + 0{,}0043\, ACT + 0{,}215\, paisfac \quad \textbf{(8.42)}$$
$$(0{,}477) \quad (0{,}130) \qquad (0{,}0155) \qquad (0{,}086)$$
$$n = 141, R^2 = 0{,}0464.$$

Não existem diferenças importantes nas estimativas MQO e MQP. A única variável explicativa significante é *paisfac*, e em ambos os casos estimamos que a probabilidade de propriedade de um computador pessoal é cerca de 0,22 maior se pelo menos um dos genitores tem curso superior.

Iniciamos revendo as propriedades dos mínimos quadrados ordinários na presença de heteroscedasticidade. Esta não causa viés ou inconsistência nos estimadores MQO, mas os erros-padrão e as estatísticas de testes usuais não serão mais válidos. Mostramos como computar erros-padrão e estatísticas t robustos em relação à heteroscedasticidade, algo que é feito rotineiramente por muitos programas econométricos. A maioria dos programas também computa uma estatística F robusta em relação à heteroscedasticidade.

Discutimos duas maneiras comuns de verificar a existência de heteroscedasticidade: o teste de Breusch-Pagan e um caso especial do teste de White. Essas duas estatísticas envolvem a regressão dos resíduos *quadrados* do MQO sobre as variáveis independentes (BP) ou sobre os valores estimados e o quadrado destes (White). Um teste F simples é assimptoticamente válido; existem também versões do multiplicador de Lagrange dos testes.

O MQO não é mais o melhor estimador linear não-viesado na presença de heteroscedasticidade. Quando a forma da heteroscedasticidade é conhecida, a estimação por mínimos quadrados generalizados (MQG) pode ser usada. Isso conduz aos mínimos quadrados ponderados como um meio de obter estimadores BLUE. As estatísticas de testes da estimação MQP são perfeitamente válidas quando o termo erro é normalmente distribuído ou assimptoticamente válido sob não normalidade. Isso supõe, é claro, que temos o modelo de heteroscedasticidade apropriado.

Mais comumente, devemos estimar um modelo quanto à heteroscedasticidade antes de aplicarmos o MQP. O estimador MQG *factível* resultante não mais será não-viesado, mas será consistente e assimptoticamente eficiente. As estatísticas habituais da regressão MQP são assimptoticamente válidas. Discutimos um método para assegurar que as variâncias estimadas sejam estritamente positivas para todas as observações, o que é necessário para aplicar o método MQP.

Como discutimos no Capítulo 7, o modelo de probabilidade linear de uma variável dependente binária terá, necessariamente, um termo erro heteroscedástico. Uma maneira simples de lidar com esse problema é computar estatísticas robustas em relação à heteroscedasticidade. Alternativamente, se todos os valores estimados (isto é, as probabilidades estimadas) estiverem estritamente entre zero e um, os mínimos quadrados ponderados poderão ser usados para a obtenção de estimadores assimptoticamente eficientes.

8.1 Quais das seguintes alternativas são conseqüências da heteroscedasticidade?

(i) Os estimadores MQO, $\hat{\beta}_j$, são inconsistentes.

(ii) A estatística F usual não mais tem uma distribuição F.

(iii) Os estimadores MQO não mais são BLUE.

8.2 Considere um modelo linear para explicar o consumo mensal de cerveja:

$$cerveja = \beta_0 + \beta_1 renda + \beta_2 preço + \beta_3 educ + \beta_4 feminino + u$$
$$\mathrm{E}(u|renda,preço,educ,feminino) = 0$$
$$\mathrm{Var}(u|renda,preço,educ,feminino) = \sigma^2 renda^2.$$

Escreva a equação transformada que tenha um termo erro homoscedástico.

8.3 Verdadeiro ou Falso: O método MQP é preferido ao MQO quando uma variável importante for omitida do modelo.

8.4 Utilizando os dados contidos no arquivo GPA3.RAW, a seguinte equação foi estimada para os alunos de uma universidade:

$$\widehat{nsgrad} = -2{,}12 + 0{,}900\ npgrad + 0{,}193\ nmgradac + 0{,}0014\ tothrs$$
$$(0{,}55)\ \ (0{,}175)\qquad (0{,}064)\qquad (0{,}0012)$$
$$[0{,}55]\ \ [0{,}166]\qquad [0{,}074]\qquad [0{,}0012]$$
$$+ 0{,}0018\ sat - 0{,}0039\ emperc + 0{,}351\ feminino - 0{,}157\ estac$$
$$(0{,}0002)\qquad (0{,}0018)\qquad (0{,}085)\qquad (0{,}098)$$
$$[0{,}0002]\qquad [0{,}0019]\qquad [0{,}079]\qquad [0{,}080]$$
$$n = 269,\ R^2 = 0{,}465.$$

Aqui, *nsgrad* é a nota obtida pelo aluno no exame final do curso, no semestre corrente, *npgrad* é uma média ponderada das notas nas diversas disciplinas cursadas no semestre, *nmgradac* é a nota do exame de final de semestre, no semestre anterior, *tothrs* é o total de créditos em horas, acumuladas até o semestre anterior, *sat* é a nota do aluno no exame de ingresso na Universidade, *emperc* é o percentil do aluno no curso médio na escola em que o aluno se formou antes de ingressar na Universidade, *feminino* é uma *dummy* de gênero e *estac* é uma *dummy* igual a um se o esporte praticado pelo aluno for praticado durante o outono. Os erros-padrão usuais e os robustos em relação à heteroscedasticidade estão registrados entre parênteses e colchetes, respectivamente.

(i) As variáveis *npgrad*, *nmgradac* e *tothrs* têm os efeitos estimados esperados? Quais dessas variáveis são estatisticamente significantes ao nível de 5%? Importa quais erros-padrão são usados?

(ii) Por que a hipótese H_0:$\beta_{npgrad} = 1$ faz sentido? Teste esta hipótese contra a alternativa bicaudal ao nível de 5%, usando ambos os erros-padrão. Descreva suas conclusões.

(iii) Verifique se existe um efeito sazonal sobre a variável *nsgrad*, usando ambos os erros-padrão. O nível de significância no qual a hipótese nula pode ser rejeitada depende do erro-padrão usado?

8.5 A variável *fuma* é uma variável binária igual a um se a pessoa fuma, e zero caso contrário. Utilizando os dados contidos no arquivo SMOKE.RAW, estimamos um modelo de probabilidade linear de *fuma*:

$$\widehat{fuma} = \underset{\substack{(0,855)\\[0,856]}}{0,656} - \underset{\substack{(0,204)\\[0,207]}}{0,069}\log(precig) + \underset{\substack{(0,026)\\[0,026]}}{0,012}\log(renda) - \underset{\substack{(0,006)\\[0,006]}}{0,029}\,educ$$

$$+ \underset{\substack{(0,006)\\[0,005]}}{0,020}\,idade - \underset{\substack{(0,00006)\\[0,00006]}}{0,00026}\,idade^2 - \underset{\substack{(0,039)\\[0,038]}}{0,101}\,restaurn - \underset{\substack{(0,052)\\[0,050]}}{0,026}\,branco$$

$$n = 807,\ R^2 = 0,062.$$

A variável *branco* é igual a um se a pessoa for branca, e zero caso contrário; as outras variáveis independentes foram definidas no Exemplo 8.7. Tanto os erros-padrão usuais como os robustos em relação à heteroscedasticidade estão informados.

(i) Existe alguma diferença importante entre os dois conjuntos de erros-padrão?

(ii) Mantendo os outros fatores fixos, se a educação aumentar em quatro anos, o que acontece com a probabilidade de fumar estimada?

(iii) Em que ponto mais um ano de idade reduz a probabilidade de fumar?

(iv) Interprete o coeficiente da variável binária *restaurn* (uma variável *dummy* igual a um se a pessoa viver em um estado em que há restrições de fumar em restaurantes).

(v) A pessoa número 206 no conjunto de dados tem as seguintes características: *precig* = 67,44, *renda* = 6.500, *educ* = 16, *idade* = 77, *restaurn* = 0, *branco* = 0 e *fuma* = 0. Calcule a probabilidade de fumar dessa pessoa e comente o resultado.

B

Fundamentos da Probabilidade

Este apêndice trata dos conceitos-chave de probabilidade básica. Os Apêndices B e C são essencialmente de recapitulação; eles não pretendem substituir um curso sobre probabilidade ou estatística. Porém, todos os conceitos sobre probabilidade e estatística que usamos neste livro são discutidos nesses apêndices.

A probabilidade por si só é de interesse dos estudiosos de negócios, economia e outras ciências sociais. Por exemplo, considere o problema de uma empresa aérea que esteja tentando decidir quantas reservas aceitar para um vôo com 100 lugares disponíveis. Se menos de 100 pessoas quiserem fazer reservas, então todas deverão ser aceitas. Mas e se mais de 100 pessoas solicitarem reserva? Uma solução segura seria aceitar no máximo 100 reservas. Porém, como algumas pessoas fazem reservas e não comparecem para o embarque, existe alguma probabilidade de que o avião não lote mesmo que sejam feitas 100 reservas. Isso resultará em perda de receita para a empresa aérea. Uma estratégia diferente seria aceitar mais de 100 reservas e esperar que algumas pessoas não compareçam para embarque, e assim o número final de passageiros seria o mais próximo possível de 100. Essa decisão corre o risco de a companhia aérea ter de compensar as pessoas que não puderam embarcar devido à venda de um número de assentos maior que o da capacidade do avião.

Uma questão natural nesse contexto é: podemos decidir sobre o número ótimo (ou o melhor) de reservas que a companhia aérea deveria fazer? Esse não é um problema trivial. Contudo, levando-se em consideração certas informações (sobre os custos da empresa aérea e a freqüência das pessoas deixarem de comparecer para o embarque), podemos usar probabilidade básica para chegar a uma solução.

B.1 VARIÁVEIS ALEATÓRIAS E SUA DISTRIBUIÇÕES DE PROBABILIDADE

Suponha que joguemos para o alto uma moeda dez vezes e contemos o número de vezes em que dê cara. Esse é um exemplo de um **experimento**. De forma geral, um experimento é qualquer procedimento que possa, pelo menos em teoria, ser repetido indefinidamente, e tem um conjunto de resultados bem definido. Poderíamos, em princípio, continuar tirando cara ou coroa repetidamente. Antes de atirarmos a moeda, sabemos que o número de caras que aparecerá será um inteiro entre 0 e 10, e, portanto, os resultados do experimento são bem definidos.

Uma **variável aleatória** é aquela que assume valores numéricos e tem um resultado que é determinado por um experimento. No exemplo da moeda, o número de caras que aparecerá em dez lances de uma moeda é um exemplo de uma variável aleatória. Antes de atirarmos a moeda dez vezes, não sabemos quantas vezes vai dar cara. Ao lançarmos a moeda dez vezes e contarmos o número de vezes que deu cara, obteremos o resultado da variável aleatória para essa particular verificação do experimento. Outra verificação poderá produzir um resultado diferente.

No exemplo das reservas da empresa aérea mencionado anteriormente, o número de pessoas que comparece para o embarque é uma variável aleatória: antes de qualquer vôo, não sabemos quantas pessoas comparecerão para embarque.

Para analisar os dados coletados em economia e nas ciências sociais, é importante ter-se um conhecimento básico das variáveis aleatórias e de suas propriedades. Seguindo as convenções tradicionais de probabilidade e estatística, ao longo dos Apêndices B e C, representaremos as variáveis aleatórias com letras maiúsculas, em geral W, X, Y e Z; os resultados particulares das variáveis aleatórias são representados pelas minúsculas correspondentes, w, x, y e z. Por exemplo, no experimento da moeda, seja X o número de vezes que apareceu cara em dez lances da moeda. Nesse caso, X não está associado com qualquer valor em particular, mas sabemos que X assumirá um valor do conjunto $\{0, 1, 2, ..., 10\}$. Um resultado particular seria, digamos, $x = 6$.

Indicamos coleções grandes de variáveis aleatórias pelo uso de subscritos. Por exemplo, se registrarmos a renda do ano passado de 20 famílias escolhidas aleatoriamente nos Estados Unidos, poderemos representar essas variáveis aleatórias por $X_1, X_2, ..., X_{20}$; os resultados particulares seriam representados por $x_1, x_2, ..., x_{20}$.

Como afirmado na definição, as variáveis aleatórias sempre são estabelecidas para assumir valores numéricos, mesmo quando descrevem eventos qualitativos. Por exemplo, considere jogar uma única moeda, na qual os dois resultados são cara e coroa. Podemos definir uma variável aleatória da seguinte forma: $X = 1$ se der cara, e $X = 0$ se der coroa.

Uma variável aleatória que somente pode assumir os valores zero e um é chamada **variável aleatória de Bernoulli** (ou **binária**). Em probabilidade básica, é tradição chamar o evento $X = 1$ de "sucesso" e o evento $X = 0$ de "fracasso". A nomenclatura sucesso-fracasso pode não corresponder à nossa noção de sucesso e fracasso em determinadas aplicações, mas é uma terminologia útil que adotaremos.

Variáveis Aleatórias Discretas

Uma **variável aleatória discreta** é a que somente assume um número finito ou infinito enumerável de valores. A noção de "infinito enumerável" significa que, embora um número infinito de valores possa ser assumido por uma variável aleatória, esses valores podem ser postos em uma correspondência um-a-um com os números inteiros positivos. Como a distinção entre "infinito enumerável" e "infinito não-enumerável" é um pouco sutil, nos concentraremos nas variáveis aleatórias discretas que assumem somente um número finito de valores. Larsen e Marx (1986, Capítulo 3) apresentam uma abordagem detalhada sobre o assunto.

Uma variável aleatória de Bernoulli é o exemplo mais simples de uma variável aleatória discreta. A única coisa que precisamos para descrever completamente o comportamento de uma variável aleatória de Bernoulli é a probabilidade que ela assume no valor um. No exemplo da moeda, se ela for "justa", então, $P(X = 1) = 1/2$ (lê-se como "a probabilidade de que X seja igual a um é de 0,5). Como a soma das probabilidades deve ser igual à unidade, $P(X = 0) = 1/2$.

Os cientistas sociais estão interessados em mais do que cara ou coroa, e, portanto, devemos considerar situações mais gerais. Novamente, considere o exemplo em que a empresa aérea tem de decidir quantas reservas aceitar para um vôo com 100 lugares disponíveis. Esse problema pode ser analisado no contexto de diversas variáveis aleatórias de Bernoulli da seguinte maneira: para um passageiro selecionado aleatoriamente, defina uma variável aleatória de Bernoulli como $X = 1$ se a pessoa aparecer para embarque, e $X = 0$ se não aparecer.

Não há nenhuma razão para pensar que a probabilidade de qualquer passageiro em particular comparecer para embarque ser 1/2; em princípio, a probabilidade pode ser qualquer número entre zero e um. Chame esse número θ, de forma que

$$P(X = 1) = \theta \quad \textbf{(B.1)}$$

$$P(X = 0) = 1 - \theta. \quad \textbf{(B.2)}$$

Por exemplo, se $\theta = 0{,}75$, existirá 75% de probabilidade de que um passageiro apareça para o embarque após ter feito a reserva e 25% de probabilidade de que o passageiro não apareça. Intuitivamente, o valor de θ é fundamental para determinar a estratégia da companhia aérea quanto à aceitação de reservas. Os métodos para estimar θ, considerando os dados históricos de reservas das companhias aéreas, são tópicos de estatística matemática, que veremos no Apêndice C.

De forma mais geral, qualquer variável aleatória discreta é completamente descrita listando seus possíveis valores e a probabilidade associada que ela assume para cada valor. Se X assumir os k possíveis valores $\{x_1, ..., x_k\}$, as probabilidades $p_1, p_2, ..., p_k$ serão definidas por

$$p_j = P(X = x_j), j = 1, 2, ..., k, \quad \textbf{(B.3)}$$

onde cada p_j estará entre zero e um e

$$p_1 + p_2 + ... + p_k = 1. \quad \textbf{(B.4)}$$

A equação (B.3) é lida como: "A probabilidade de X assumir o valor x_j é igual a p_j".

As equações (B.1) e (B.2) mostram que as probabilidades de sucesso e fracasso de uma variável aleatória de Bernoulli são determinadas inteiramente pelo valor de θ. Como as variáveis aleatórias de Bernoulli são tão freqüentes, temos uma notação especial para elas: $X \sim$ Bernoulli(θ) é lida como "X tem uma distribuição de Bernoulli com probabilidade de sucesso igual a θ."

A **função densidade de probabilidade** (**fdp**) de X resume as informações relativas aos possíveis resultados de X e as probabilidades correspondentes:

$$f(x_j) = p_j, j = 1, 2, ..., k, \quad \textbf{(B.5)}$$

com $f(x) = 0$ de qualquer x não igual a x_j para algum j. Em outras palavras, para qualquer número real x, $f(x)$ será a probabilidade que a variável aleatória X assumirá para o valor particular de x. Quando lidamos com mais de uma variável aleatória, algumas vezes é útil subscrever a fdp em questão: f_X é a fdp de X, f_Y é a fdp de Y, e assim por diante.

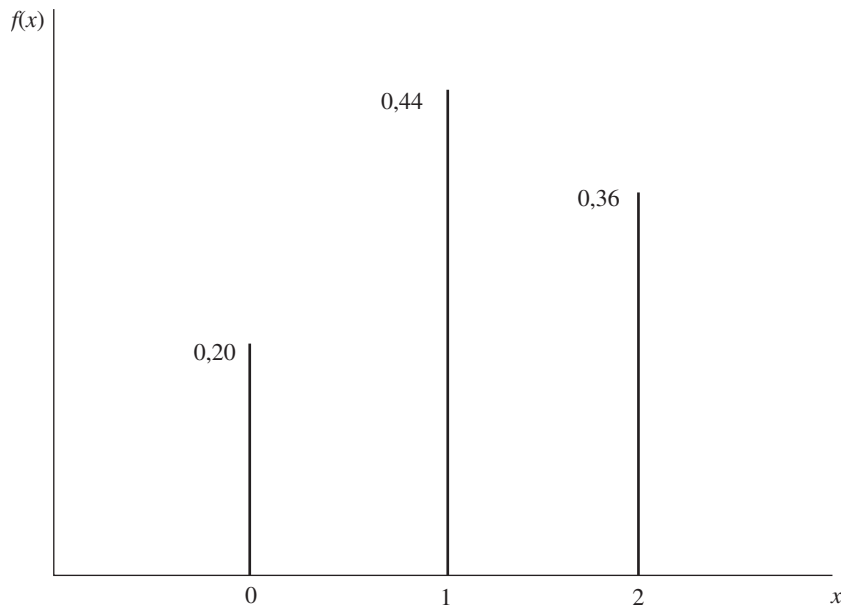
Dada a fdp de qualquer variável aleatória discreta, é simples calcular a probabilidade de qualquer evento envolvendo aquela variável aleatória. Por exemplo, suponha que X seja o número de pontos feitos por um jogador de basquetebol a cada dois lances livres, de forma que X pode assumir os três valores $\{0,1,2\}$. Assuma que a fdp de X seja dada por

$$f(0) = 0{,}20, f(1) = 0{,}44, \text{ e } f(2) = 0{,}36.$$

A soma das três probabilidades é igual a um, como deveria ser. Usando essa fdp, podemos calcular a probabilidade de que o jogador converta *pelo menos* um lance livre: $P(X \geq 1) = P(X = 1) + P(X = 2) = 0{,}44 + 0{,}36 = 0{,}80$. A fdp de X é mostrada na Figura B.1.

Figura B.1

A fdp do número de lances livres convertidos a cada duas tentativas.

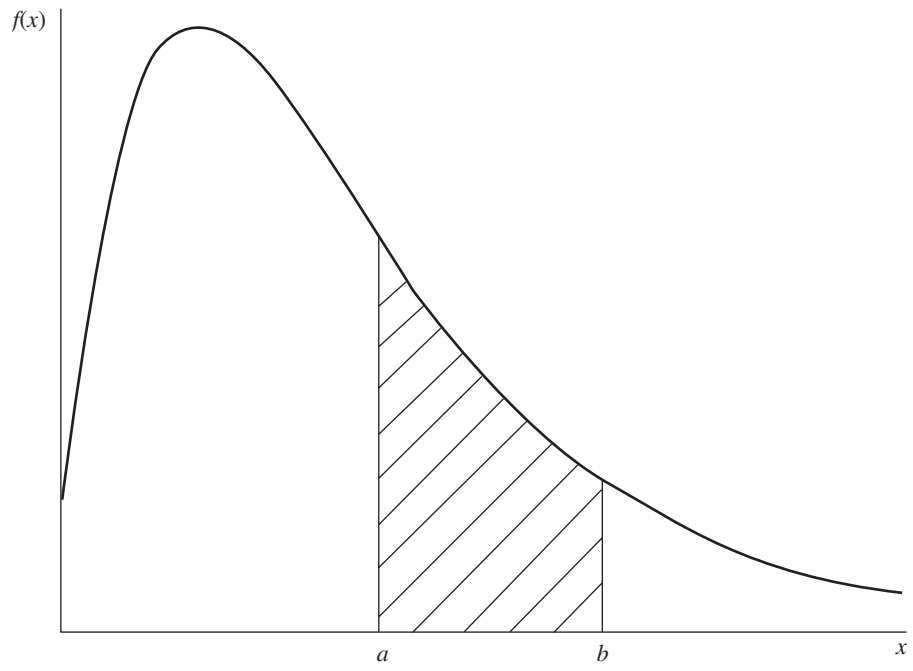


Variáveis Aleatórias Contínuas

Uma variável X será uma **variável aleatória contínua** se assumir qualquer valor real com probabilidade zero. Essa definição é um tanto quanto contra-intuitiva, já que em qualquer aplicação acabaremos observando algum resultado para uma variável aleatória. A idéia é que uma variável aleatória contínua X pode assumir tantos valores possíveis que não podemos enumerá-los ou compará-los com os inteiros positivos, de modo que a consistência lógica garante que X pode assumir cada valor com probabilidade zero. Embora as medidas sejam sempre discretas na prática, as variáveis aleatórias que assumem numerosos valores são melhor tratadas como contínuas. Por exemplo, a medida mais refinada do preço de um bem é em termos de centavos. Podemos nos imaginar relacionando todos os possíveis valores de preços ordenadamente (mesmo que a lista possa continuar indefinidamente), o que tecnicamente faz com que preço seja uma variável aleatória discreta. Porém, existem tantos valores possíveis de preços que o uso da mecânica das variáveis aleatórias discretas não é viável.

Podemos definir uma função densidade de probabilidade para variáveis aleatórias contínuas, e, como acontece com as variáveis aleatórias discretas, a fdp fornecerá informações sobre os prováveis resultados da variável aleatória. Porém, como também não faz sentido discutir a probabilidade de que uma variável aleatória contínua assuma um valor em particular, usamos a fdp de uma variável aleatória contínua somente para computar eventos envolvendo uma diversidade de valores. Por exemplo, se a e b forem constantes onde $a < b$, a probabilidade de X estar entre os números a e b, $P(a \leq X \leq b)$, será a *área* sob a fdp entre os pontos a e b, como mostrado na Figura B.2. Se você estiver familiarizado com cálculo diferencial, você reconhecerá isso como a *integral* da função f entre os pontos a e b. A área total sob a fdp deve sempre ser igual a um.

Figura B.2
A probabilidade que X esteja entre os pontos *a* e *b*.



Ao computar probabilidades para variáveis aleatórias contínuas, é mais fácil trabalhar com a **função de distribuição cumulativa** (**fdc**). Se X for qualquer variável aleatória, então, sua fdc será definida por qualquer numero x real pela equação

$$F(x) \equiv \mathrm{P}(X \leq x). \qquad \textbf{(B.6)}$$

Para variáveis aleatórias discretas, (B.6) será obtida somando a fdp para todos os valores x_j tais que $x_j \leq x$. Para uma variável aleatória contínua, $F(x)$ será a área sob a fdp, f, à esquerda do ponto x. Como $F(x)$ é simplesmente uma probabilidade, ela estará sempre entre 0 e 1. Além disso, se $x_1 < x_2$, então, $\mathrm{P}(X \leq x_1) \leq \mathrm{P}(X \leq x_2)$, isto é, $F(x_1) \leq F(x_2)$. Isso significa que uma fdc é uma função crescente (ou pelo menos não-decrescente) de x.

Duas propriedades importantes das fdcs que são úteis no cálculo de probabilidades são as seguintes:

Para qualquer número c, $\mathrm{P}(X > c) = 1 - F(c)$. **(B.7)**

Para quaisquer números $a < b$, $\mathrm{P}(a < X \leq b) = F(b) - F(a)$. **(B.8)**

Em nosso estudo da econometria, usaremos as fdcs para calcular probabilidades somente de variáveis aleatórias contínuas, caso em que não importa se as desigualdades nas especificações probabilísticas são estritas ou não. Ou seja, para uma variável aleatória contínua X,

$$P(X \geq c) = P(X > c) \tag{B.9}$$

e

$$P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b). \tag{B.10}$$

Combinadas com (B.7) e (B.8), as equações (B.9) e (B.10) expandem bastante os cálculos de probabilidade que podem ser feitos com o uso de fdcs contínuas.

As funções de distribuições cumulativas foram tabuladas para todas as distribuições contínuas importantes em probabilidade e estatística. A mais conhecida delas é a distribuição normal, da qual trataremos com algumas distribuições relacionadas na Seção B.5

B.2 DISTRIBUIÇÕES CONJUNTAS, DISTRIBUIÇÕES CONDICIONAIS E INDEPENDÊNCIA

Em economia, geralmente estamos interessados na ocorrência de eventos que envolvem mais de uma variável aleatória. No exemplo das reservas da companhia aérea anteriormente referido, esta pode estar interessada na probabilidade de que uma pessoa que faz uma reserva compareça para embarque *e* que seja uma pessoa que viaje a negócios; esse é um exemplo de uma *probabilidade conjunta.* Ou a empresa aérea pode estar interessada na seguinte *probabilidade condicional*: condicional à pessoa ser uma pessoa que viaje a negócios, qual é a probabilidade de que ela compareça para embarque? Nas próximas duas subseções, formalizaremos a noção de distribuições conjunta e condicional e a importante noção de *independência* das variáveis aleatórias.

Distribuições Conjuntas e Independência

Sejam X e Y variáveis aleatórias discretas. Então, (X,Y) têm uma **distribuição conjunta**, que é totalmente definida pela *função densidade de probabilidade conjunta* de (X,Y):

$$f_{X,Y}(x,y) = P(X = x, Y = y), \tag{B.11}$$

onde o lado direito é a probabilidade de que $X = x$ e $Y = y$. Quando X e Y são contínuas, uma fdp conjunta também pode ser definida, mas não trataremos de tais detalhes, pois fdps conjuntas de variáveis aleatórias contínuas não são explicitamente usadas neste livro.

Em um caso, é fácil obter a fdp conjunta se forem dadas as fdps de X e Y. Em particular, as variáveis aleatórias X e Y são independentes se, e somente se,

$$f_{X,Y}(x,y) = f_X(x) f_y(y) \tag{B.12}$$

para todos os x e y, quando f_X for a fdp de X e f_Y for a fdp de Y. No contexto de mais de uma variável aleatória, as fdps f_X e f_Y são freqüentemente chamadas *funções de densidade de probabilidade marginal* para distingui-las da fdp conjunta $f_{X,Y}$. Essa definição de independência é válida para variáveis aleatórias discretas e contínuas.

Para entendermos o significado de (B.12) é mais fácil lidar com o caso discreto. Se X e Y forem discretas, então, (B.12) será a mesma coisa que

$$P(X = x,Y = y) = P(X = x)P(Y = y); \qquad \textbf{(B.13)}$$

em outras palavras, a probabilidade de que $X = x$ e $Y = y$ é o produto das duas probabilidades $P(X = x)$ e $P(Y = y)$. Uma implicação de (B.13) é que as probabilidades conjuntas são razoavelmente fáceis de serem calculadas, já que elas apenas exigem o conhecimento de $P(X = x)$ e $P(Y = y)$.

Se as variáveis aleatórias não forem independentes, então, elas são *dependentes*.

EXEMPLO B.1

(Arremessos de Lances Livres)

Considere um jogador de basquetebol fazendo dois lances livres. Seja X uma variável aleatória de Bernoulli igual a um se ele converter o primeiro arremesso, e zero, caso contrário. Seja Y uma variável aleatória de Bernoulli igual a um se ele converter o segundo arremesso. Suponha que ele seja um jogador que converte 80% dos arremessos, de forma que $P(X = 1) = P(Y = 1) = 0,8$. Qual é a probabilidade de o jogador converter os dois arremessos?

Se X e Y forem independentes, podemos facilmente responder a essa pergunta: $P(X = 1, Y = 1) = P(X = 1)(P(Y = 1) = (0,8)(0,8) = 0,64$. Portanto, existe 64% de probabilidade de converter ambos os lances livres. Se a probabilidade de converter o segundo arremesso depender de o primeiro arremesso ter sido convertido — isto é, X e Y não são independentes — então, esse cálculo simples não será válido.

A independência de variáveis aleatórias é um conceito muito importante. Na próxima subseção, mostraremos que, se X e Y forem independentes, conhecer resultado de X não altera as probabilidades dos possíveis resultados de Y, e vice-versa. Um fato útil sobre a questão da independência é que se X e Y forem independentes e definirmos novas variáveis aleatórias $g(X)$ e $h(Y)$ para quaisquer funções g e h, então, essas novas variáveis aleatórias também serão independentes.

Não há necessidade de parar em duas variáveis aleatórias. Se $X_1, X_2, ..., X_n$ forem variáveis aleatórias discretas, então, suas fdps conjuntas serão $f(x_1, x_2, ..., x_n) = P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$. As variáveis aleatórias $X_1, X_2, ..., X_n$ serão **variáveis aleatórias independentes** se, e somente se, suas fdps conjuntas forem o produto das fdps individuais para quaisquer $(x_1, x_2, ..., x_n)$. Essa definição de independência também é válida para variáveis aleatórias contínuas.

A noção de independência desempenha um papel importante na obtenção de algumas distribuições clássicas em probabilidade e estatística. Anteriormente, definimos uma variável aleatória de Bernoulli como uma variável aleatória zero-um indicando se ocorre algum evento ou não. Freqüentemente, estamos interessados no número de sucessos em uma seqüência de ensaios de Bernoulli *independentes*.

Um exemplo padrão de ensaios de Bernoulli independentes é jogar repetidamente uma moeda. Como o resultado de qualquer lance particular não tem nada a ver com os resultados dos outros lances, a independência é uma hipótese apropriada.

A independência é muitas vezes uma aproximação razoável em situações mais complicadas. No exemplo das reservas da companhia aérea, suponha que a companhia aceite n reservas para determinado vôo. De cada $i = 1, 2, ..., n$, seja Y a variável aleatória de Bernoulli indicando se o passageiro i aparece para embarque: $Y_i = 1$ se o passageiro i aparecer para embarque, e $Y_i = 0$, caso contrário. Definindo θ novamente como a probabilidade de sucesso (usando as reservas), cada Y_i terá uma distribuição de Bernoulli (θ). Como uma aproximação, podemos assumir que os Y_i são independentes entre si, embora isso não seja exatamente verdadeiro na realidade: algumas pessoas viajam em grupo, o que significa que se uma pessoa comparecerá ou não para embarque não é verdadeiramente independente de se as outras pessoas comparecerão ou não. Porém, modelar esse tipo de dependência é complexo, de modo que podemos querer usar a independência como uma aproximação.

A variável de interesse principal é o número total de passageiros que comparecem para embarque das n reservas; chamemos essa variável de X. Como cada Y_i será igual à unidade quando uma pessoa comparece para embarque, podemos escrever $X = Y_1 + Y_2 + ... + Y_n$. Agora, assumindo que cada Y_i tem probabilidade de sucesso θ e que os Y_i são independentes, é possível mostrar que X tem uma **distribuição binomial**. Isto é, a função densidade de probabilidade de X é

$$f(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x},\ x = 0, 1, 2, ..., n, \qquad \textbf{(B.14)}$$

onde $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, e para qualquer inteiro n, $n!$ (lê-se "fatorial de n") é definido como $n! = n \cdot (n-1)(n-2)...1$. Por convenção, $0! = 1$. Quando uma variável aleatória X tem a fdp dada em (B.14), escrevemos $X \sim \text{Binomial}(n,\theta)$. A equação (B.14) pode ser usada para calcular $P(X = x)$ para qualquer valor de x de 0 a n.

Se o vôo tiver 100 lugares disponíveis, a empresa aérea estará interessada em $P(X > 100)$. Suponha, inicialmente, que $n = 120$, de modo que a companhia aérea aceitará 120 reservas, e que a probabilidade de que cada pessoa compareça para embarque seja $\theta = 0{,}85$. Então, $P(X > 100) = P(X = 101) + P(X = 102) + ... + P(X = 120)$, e cada uma das probabilidades na soma poderá ser encontrada pela equação (B.14) com $n = 120$, $\theta = 0{,}85$ e o valor apropriado de x (101 a 120). Esse é um cálculo difícil de ser feito manualmente, mas muitos programas estatísticos possuem comandos para computar esse tipo de probabilidade. Nesse caso, a probabilidade de que mais de 100 pessoas comparecerão para embarque é cerca de 0,659, o que provavelmente é um risco de excesso de reservas maior do que a companhia aérea deseja tolerar. Se, em vez disso, o número de reservas for 110, a probabilidade de que mais de 100 pessoas comparecerão para embarque será de apenas 0,024.

Distribuições Condicionais

Em econometria, geralmente estamos interessados em como uma variável aleatória, vamos chamá-la Y, está relacionada com uma ou mais das outras variáveis. Por enquanto, suponha que haja somente uma variável em cujos efeitos estamos interessados, vamos chamá-la X. O máximo que podemos saber sobre como X afeta Y está contido na **distribuição condicional** da Y, dado X. Essa informação é resumida pela *função de densidade de probabilidade condicional*, definida por

$$f_{Y|X}(y|x) = f_{X,Y}(x,y)/f_X(x) \qquad \textbf{(B.15)}$$

para todos os valores de x de tal forma que $f_X(x) > 0$. A interpretação de (B.15) é mais fácil de ser vista quando X e Y são discretas. Então,

$$f_{Y|X}(y|x) = \mathrm{P}(Y = y|X = x), \tag{B.16}$$

onde o lado direito é lido como "a probabilidade de $Y = y$ em decorrência de $X = x$." Quando Y é contínuo, $f_{Y|X}(y|x)$ não é interpretada diretamente como uma probabilidade, pelas razões explicadas anteriormente, mas as probabilidades condicionais são encontradas computando áreas sob a fdp condicional.

Uma característica importante das distribuições condicionais é que, se X e Y forem variáveis aleatórias independentes, o conhecimento dos valores assumidos por X não nos diz nada sobre a probabilidade de que Y assuma diversos valores (e vice-versa). Isto é, $f_{Y|X}(y|x) = f_Y(y)$ e $f_{X|Y}(x|y) = f_X(x)$.

EXEMPLO B.2

(Arremessos de Lances Livres)

Considere novamente o exemplo dos lances livres no basquetebol, quando dois lances livres devem ser tentados. Assuma que a densidade condicional seja

$$f_{Y|X}(1|1) = 0{,}85, f_{Y|X}(0|1) = 0{,}15$$
$$f_{Y|X}(1|0) = 0{,}70, f_{Y|X}(0|0) = 0{,}30.$$

Isso significa que a probabilidade de o jogador converter o segundo lance depende de o primeiro lance ter sido convertido: se o primeiro lance foi convertido, a probabilidade de converter o segundo lance é de 0,85; se o primeiro lance foi perdido, a probabilidade de converter o segundo lance é de 0,70. Isso implica que X e Y não são independentes; eles são dependentes.

Ainda podemos computar $\mathrm{P}(X = 1, Y = 1)$ desde que conheçamos $\mathrm{P}(X = 1)$. Assuma que a probabilidade de converter o primeiro lance livre seja 0,8, isto é, $\mathrm{P}(X = 1) = 0{,}8$. Então, de (B.15), teremos

$$\mathrm{P}(X = 1, Y = 1) = \mathrm{P}(Y = 1|X = 1)\cdot\mathrm{P}(X = 1) = (0{,}85)(0{,}8) = 0{,}68.$$

B.3 CARACTERÍSTICAS DAS DISTRIBUIÇÕES DE PROBABILIDADE

Para muitos propósitos, estaremos interessados em somente alguns poucos aspectos das distribuições das variáveis aleatórias. As características de interesse podem ser classificadas em três categorias: medidas de tendência central, medidas de variabilidade ou intervalo e medidas de associação entre duas variáveis aleatórias. Trataremos desta última na Seção B.4.

Uma Medida de Tendência Central: O Valor Esperado

O valor esperado é um dos mais importantes conceitos da probabilidade que encontraremos em nosso estudo da econometria. Se X for uma variável aleatória, o **valor esperado** (ou esperança) de X, representado por $\mathrm{E}(X)$ e algumas vezes por μ_X, ou simplesmente μ, é uma média ponderada de todos os possíveis

valores de X. Os pesos são determinados pela função de densidade de probabilidade. Algumas vezes, o valor esperado é chamado *média populacional*, especialmente quando queremos enfatizar que X representa alguma variável em uma população.

A definição precisa do valor esperado é mais simples no caso em que X é uma variável aleatória discreta assumindo um número finito de valores, digamos, $\{x_1, ..., x_k\}$. Seja $f(x)$ a função de densidade de probabilidade de X. O valor esperado de X será a média ponderada

$$\mathrm{E}(X) = x_1 f(x_1) + x_2 f(x_2) + ... + x_k f(x_k) \equiv \sum_{j=1}^{k} x_j f(x_j) \qquad \textbf{(B.17)}$$

Essa expressão é facilmente calculada dados os valores da fdp de cada possível resultado de X.

EXEMPLO B.3

(Calculando um Valor Esperado)

Suponha que X assuma os valores -1, 0 e 2 com probabilidades 1/8, 1/2 e 3/8, respectivamente. Então,

$$\mathrm{E}(X) = (-1)\cdot(1/8) + 0\cdot(1/2) + 2\cdot(3/8) = 5/8.$$

Esse exemplo ilustra uma coisa curiosa sobre os valores esperados: o valor esperado de X pode ser um número que não é sequer um possível resultado de X. Sabemos que X assume os valores -1, 0 e 2, ainda que seu valor esperado seja 5/8. Isso torna o valor esperado deficiente para resumir a tendência central de certas variáveis aleatórias discretas, mas cálculos como o que acabamos de mostrar podem ser úteis, como veremos mais tarde.

Se X for uma variável aleatória contínua, então, $\mathrm{E}(X)$ será definido como uma integral:

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f(x)dx, \qquad \textbf{(B.18)}$$

que assumimos como bem definida. Isso ainda pode ser interpretado como uma média ponderada. Para as distribuições contínuas mais comuns, $\mathrm{E}(X)$ é um número que é um possível resultado de X. Neste livro, não precisaremos calcular valores esperados usando integração, embora utilizemos de alguns resultados bem conhecidos de probabilidade de valores esperados de variáveis aleatórias especiais.

Dada uma variável aleatória X e uma função $g(\cdot)$, podemos criar uma nova variável aleatória $g(X)$. Por exemplo, se X for uma variável aleatória, então, X^2 e $\log(X)$ (se $X > 0$) também serão variáveis aleatórias. O valor esperado de $g(X)$ será, de novo, simplesmente uma média ponderada:

$$\mathrm{E}[g(X)] = \sum_{j=1}^{k} g(x_j) f_X(x_j) \qquad \textbf{(B.19)}$$

ou, para uma variável aleatória contínua,

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)dx. \qquad \textbf{(B.20)}$$

EXEMPLO B.4

(Valor Esperado de X^2)

Para a variável aleatória no Exemplo B.3, seja $g(X) = X^2$. Então,

$$\mathrm{E}(X^2) = (-1)^2(1/8) + (0)^2(1/2) + (2)^2(3/8) = 13/8.$$

No Exemplo B.3, calculamos $\mathrm{E}(X) = 5/8$, de forma que $[\mathrm{E}(X)]^2 = 25/64$. Isso mostra que $\mathrm{E}(X^2)$ *não* é o mesmo que $[\mathrm{E}(X)]^2$. De fato, para uma função não-linear $g(X)$, $\mathrm{E}[g(X)] \neq g[\mathrm{E}(X)]$ (exceto em casos muito especiais).

Se X e Y forem variáveis aleatórias, então, $g(X,Y)$ será uma variável aleatória para qualquer função g, e assim poderemos definir sua esperança. Quando X e Y são ambas discretas, assumindo os valores $\{x_1, x_2, ..., x_k\}$ e $\{y_1, y_2, ..., y_m\}$, respectivamente, o valor esperado será

$$\mathrm{E}[g(X, Y)] = \sum_{h=1}^{k} \sum_{j=1}^{m} g(x_h,y_j)\, f_{X,Y}(x_h,y_j),$$

onde $f_{X,Y}$ será a fdp conjunta de (X,Y). A definição é mais complicada para variáveis aleatórias contínuas, pois envolve integração; não precisamos dela aqui. A extensão para mais de duas variáveis aleatórias é fácil de ser feita.

Propriedades dos Valores Esperados

Em econometria, não nos preocupamos muito em calcular os valores esperados de diversas distribuições; os cálculos principais já foram feitos muitas vezes, e em grande parte os aceitaremos sem questionar. Teremos que manipular alguns valores esperados usando umas poucas regras simples. Elas são tão importantes que as rotulamos:

PROPRIEDADE E.1
Para qualquer constante c, $\mathrm{E}(c) = c$.

PROPRIEDADE E.2
Para quaisquer constantes a e b, $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$.

Uma implicação útil de E.2 é que, se $\mu = \mathrm{E}(X)$, e definirmos uma nova variável aleatória como $Y = X - \mu$, então, $\mathrm{E}(Y) = 0$; em E.2, considere $a = 1$ e $b = -\mu$.

Como um exemplo da propriedade E.2, seja X a temperatura medida em graus Celsius, ao meio dia de determinado dia, em determinada localidade; suponha que a temperatura esperada seja $\mathrm{E}(X) = 25$. Se Y for a temperatura medida em graus Fahrenheit, então, $Y = 32 + (9/5)X$. Pela propriedade E.2, a temperatura esperada em Fahrenheit será $\mathrm{E}(Y) = 32 + (9/5)\cdot\mathrm{E}(X) = 32 + (9/5)\cdot 25 = 77$.

De forma geral, é fácil calcular o valor esperado de uma função linear de diversas variáveis aleatórias.

PROPRIEDADE E.3

Se $\{a_1, a_2, ..., a_n\}$ forem constantes e $\{X_1, X_2, ..., X_n\}$ forem variáveis aleatórias, então,

$$E(a_1X_1 + a_2X_2 + ... + a_nX_n) = a_1E(X_1) + a_2E(X_2) + ... + a_nE(X_n).$$

Ou, usando a notação de somatórios,

$$E(\sum_{i=1}^{n} a_iX_i) = \sum_{i=1}^{n} a_iE(X_i). \qquad \textbf{(B.21)}$$

Como um caso especial dessa equação, temos (com cada $a_1 = 1$)

$$E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i), \qquad \textbf{(B.22)}$$

de forma que o valor esperado da soma será a soma dos valores esperados. Essa propriedade é usada com freqüência para derivações em estatística matemática.

EXEMPLO B.5

(Encontrando a Receita Esperada)

Sejam X_1, X_2 e X_3 os números de pizzas pequenas, médias e grandes, respectivamente, vendidas durante o dia em uma pizzaria. Elas são variáveis aleatórias com valores esperados $E(X_1) = 25$, $E(X_2) = 57$ e $E(X_3) = 40$. Os preços das pizzas pequena, média e grande são 5,50, 7,60 e 9,15 (em dólares). Portanto, a receita esperada das vendas de pizzas em determinado dia será

$$\begin{aligned} E(5{,}50\,X_1 + 7{,}60\,X_2 + 9{,}15\,X_3) &= 5{,}50\,E(X_1) + 7{,}60\,E(X_2) + 9{,}15\,E(X_3) \\ &= 5{,}50(25) + 7{,}60(57) + 9{,}15(40) = 936{,}70, \end{aligned}$$

isto é, 936,70 dólares. A receita efetiva de qualquer dia particular geralmente será diferente desse valor, mas essa é a receita *esperada*.

Também podemos usar a Propriedade E.3 para mostrar que se $X \sim$ Binomial(n,θ), então, $E(X) = n\theta$. Ou seja, o número esperado de sucessos em n ensaios de Bernoulli é simplesmente o número de ensaios vezes a probabilidade de sucesso de qualquer ensaio particular. Isso será facilmente observado escrevendo X como $X = Y_1 + Y_2 + ... Y_n$, onde cada $Y_i \sim$ Bernoulli(θ). Então,

$$E(X) = \sum_{i=1}^{n} E(Y_i) = \sum_{i=1}^{n} \theta = n\theta.$$

Podemos aplicar esse resultado no exemplo das reservas da companhia aérea, quando ela aceita $n = 120$ reservas, e a probabilidade de comparecimento para embarque é $\theta = 0{,}85$. O número *esperado* de pessoas comparecendo para embarque é $120(0{,}85) = 102$. Portanto, se há 100 lugares disponíveis,

o número esperado de pessoas que comparecerão para embarque é grande demais; isso terá alguma influência na conclusão de ser uma boa idéia a companhia aceitar 120 reservas.

Na realidade, o que a companhia aérea poderia fazer seria definir uma função do lucro que considerasse a receita ganha por lugar vendido e o custo por passageiro que seja impedido de embarcar. Essa função do lucro será aleatória, pois o número efetivo de pessoas que comparecerão para embarque é aleatório. Seja r a receita líquida correspondente a cada passageiro. (Para simplificar, você pode pensar nisso como sendo o preço da passagem). Seja i a indenização devida a cada passageiro que não puder embarcar. Nem r nem i são aleatórios; eles são assumidos como conhecidos pela companhia aérea. Seja Y o lucro do vôo. Então, com 100 lugares disponíveis,

$$\begin{aligned} Y &= rX \text{ se } X \leq 100 \\ &= 100r - i(X - 100) \text{ se } X > 100. \end{aligned}$$

A primeira equação mostra o lucro se não mais que 100 pessoas comparecerem para embarque; a segunda equação é o lucro se mais de 100 pessoas comparecerem para embarque. (Nesse último caso, a receita líquida da venda de passagens é $100r$, pois foram vendidos todos os 100 lugares, e, então, $i(X - 100)$ é o custo de aceitar mais de 100 reservas). Usando o fato de que X tem uma distribuição Binomial $(n,0,85)$, onde n é o número de reservas feitas, os lucros esperados, $E(Y)$ poderão ser encontrados como uma função de n (e r e i). Calcular $E(Y)$ diretamente seria muito difícil, mas poderá ser encontrado rapidamente usando-se um computador. Uma vez os valores de r e i tenham sido dados, o valor de n que maximiza os lucros esperados poderá ser encontrado pesquisando-se diferentes valores de n.

Outra Medida de Tendência Central: A Mediana

O valor esperado é somente uma possibilidade para definir a tendência central de uma variável aleatória. Outra medida de tendência central é a **mediana**. Uma definição geral de mediana é complicada demais para nosso propósito. Se X for uma variável contínua, então, a mediana de X, digamos m, será um valor tal que metade da área de uma fdp está à esquerda de m, e a outra metade está à direita de m.

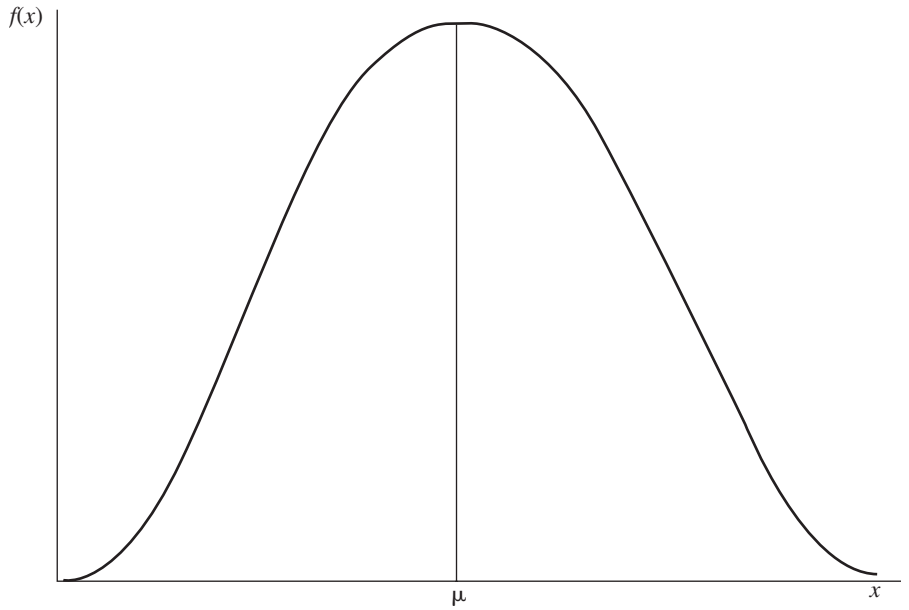
Quando X for uma variável discreta e assumir um número ímpar finito de valores, a mediada será obtida ordenando-se os possíveis valores de X e então selecionando-se o valor que estiver no centro. Por exemplo, se X assumir os valores $\{-4,0,2,8,10,13,17\}$, então, o valor mediano de X será 8. Se X assumir um número par de valores, existirão, na realidade, dois valores medianos; algumas vezes calcula-se a média desses números para obter um único valor mediano. Assim, se X assumir os valores $\{-5,3,9,17\}$, os valores medianos serão 3 e 9; se calcularmos a média desses números obteremos uma mediana igual a 6.

Em geral, a mediana, algumas vezes indicada por $\text{Med}(X)$, e o valor esperado $E(X)$, são diferentes. Nenhum é "melhor" que o outro como uma medida de tendência central; ambos são maneiras válidas de indicar o centro da distribuição de X. Em um caso especial, a mediana e o valor esperado (ou média) são os mesmos. Se X tiver uma **distribuição simétrica** em torno do valor μ, então, μ será tanto o valor esperado como a mediana. Matematicamente, a condição será $f(\mu + x) = f(\mu - x)$ para todo x. Esse caso está ilustrado na Figura B.3.

Medidas de Variabilidade: Variância e Desvio-Padrão

Embora a tendência central de uma variável aleatória seja valiosa, ela não nos diz tudo que queremos saber sobre a distribuição de uma variável aleatória. A Figura B.4 mostra as fdp de duas variáveis aleatórias com a mesma média. Claramente, a distribuição de X é mais concentrada em relação à sua média que a distribuição de Y. Gostaríamos de ter uma maneira simples de resumir isso.

Figura B.3
Uma distribuição de probabilidade simétrica.



Variância

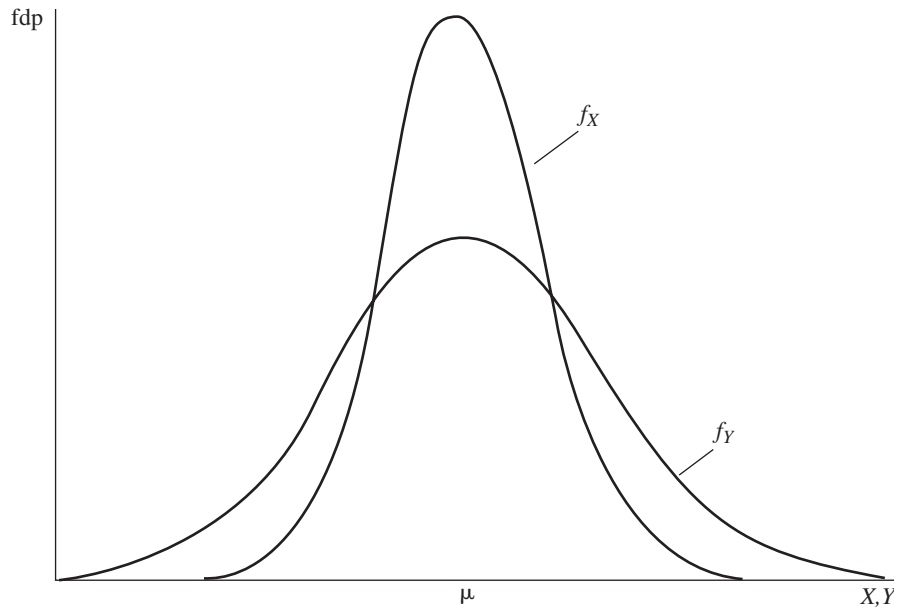
Para uma variável aleatória X, seja $\mu = \mathrm{E}(X)$. Há várias maneiras de medir o quanto X está distante de seu valor esperado, mas a mais simples de trabalhar algebricamente é a diferença elevada ao quadrado, $(X - \mu)^2$. (A elevação ao quadrado serve para eliminar o sinal da medida da distância; o valor positivo resultante corresponde à nossa noção intuitiva de distância.) Essa distância em si é uma variável aleatória, já que ela pode mudar a cada resultado de X. Da mesma forma que precisamos de um número para resumir a tendência central de X, precisamos de um número que nos informe o quanto X está distante de μ, *em média*. Um desses números é a **variância**, que nos informa a distância esperada de X até sua média:

$$\mathrm{Var}(X) \equiv \mathrm{E}(X - \mu)^2]. \qquad \textbf{(B.23)}$$

A variância é algumas vezes representada por σ_x^2, ou simplesmente σ^2, quando o contexto é claro. De (B.3), deduz-se que a variância é sempre não-negativa.

Como um instrumento computacional, é interessante observar que

Figura B.4
Variáveis aleatórias com a mesma média, mas com distribuições diferentes.



$$\sigma^2 = \mathrm{E}(X^2 - 2X\mu + \mu^2) = \mathrm{E}(X^2) - 2\mu^2 + \mu^2 = \mathrm{E}(X^2) - \mu^2. \quad \textbf{(B.24)}$$

Usando (B.23) ou (B.24), não precisamos fazer a distinção entre variáveis aleatórias discretas e contínuas: a definição de variância é a mesma em qualquer dos casos. Na maioria da vezes, primeiro calculamos $\mathrm{E}(X)$, depois $\mathrm{E}(X^2)$, e, então, usamos a fórmula de (B.4). Por exemplo, se $X \sim$ Bernoulli(θ), então, $\mathrm{E}(X) = \theta$, e como $X^2 = X$, $\mathrm{E}(X^2) = \theta$. Deduz-se da equação (B.24) que $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2 = \theta - \theta^2 = \theta(1 - \theta)$.

São apresentadas a seguir duas importantes propriedades da variância.

PROPRIEDADE VAR.1

$\mathrm{Var}(X) = 0$ se, e somente se, houver uma constante c, de tal forma que $\mathrm{P}(X = c) = 1$, em cujo caso, $\mathrm{E}(X) = c$.

Essa primeira propriedade diz que a variância de qualquer constante é zero, e se uma variável aleatória tiver variância zero, então, ela será essencialmente constante.

PROPRIEDADE VAR.2

Para quaisquer constantes a e b, $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.

Isso significa que a adição de uma constante a uma variável aleatória não altera a variância, mas a multiplicação de uma variável aleatória por uma constante aumenta a variância por um fator igual ao *quadrado* daquela constante. Por exemplo, se X representar a temperatura em graus Celsius e $Y = 32 + (9/5)X$ for a temperatura em graus Fahrenheit, então, $\text{Var}(Y) = (9/5)^2\text{Var}(X) = (81/25)\text{Var}(X)$.

Desvio-Padrão

O **desvio-padrão** de uma variável aleatória, representado por $\text{dp}(X)$, é simplesmente a raiz quadrada positiva da variância: $\text{dp}(X) \equiv +\sqrt{\text{Var}(X)}$. O desvio-padrão algumas vezes é representado por σ_x, ou simplesmente σ, quando a variável aleatória é entendida. Duas propriedades do desvio-padrão resultam das propriedades VAR.1 e VAR.2.

PROPRIEDADE DP.1

Para qualquer constante c, $\text{dp}(c) = 0$.

PROPRIEDADE DP.2

Para quaisquer constantes a e b,

$$\text{dp}(aX + b) = |a|\text{dp}(X).$$

Em particular, se $a > 0$, então, $\text{dp}(aX) = a \cdot \text{dp}(X)$.

Essa última propriedade faz com que seja mais natural trabalhar com o desvio-padrão do que com a variância. Por exemplo, suponha que X seja uma variável aleatória medida em milhares de dólares, digamos renda. Se definirmos $Y = 1.000X$, então, Y será a renda medida em dólares. Suponha que $\text{E}(X) = 20$ e $\text{dp}(X) = 6$. Então, $\text{E}(Y) = 1{,}000\text{E}(X) = 20.000$ e $\text{dp}(Y) = 1.000 \cdot \text{dp}(X) = 6.000$, de forma que o valor esperado e o desvio-padrão crescem pelo mesmo fator, 1.000. Se tivéssemos trabalhado com a variância, teríamos $\text{Var}(Y) = (1.000)^2\text{Var}(X)$, de forma que a variância de Y é um milhão de vezes maior que a variância de X.

Padronizando uma Variável Aleatória

Como uma aplicação das propriedades da variância e do desvio-padrão — e um tópico de interesse prático por si mesmo — suponha que, dada uma variável aleatória X, definamos uma nova variável aleatória subtraindo sua média μ e dividindo o resultado por seu desvio-padrão σ:

$$Z \equiv \frac{X - \mu}{\sigma}, \qquad \textbf{(B.25)}$$

que podemos escrever como $Z = aX + b$, onde $a \equiv (1/\sigma)$ e $b \equiv -(\mu/\sigma)$. Então, de acordo com a propriedade E.2,

$$\text{E}(Z) = a\text{E}(X) + b = (\mu/\sigma) - (\mu/\sigma) = 0.$$

Da propriedade Var.2,

$$\text{Var}(Z) = a^2\text{Var}(X) = (\sigma^2/\sigma^2) = 1.$$

Portanto, a variável aleatória Z tem uma média zero e uma variância (e portanto um desvio-padrão) igual a um. Esse procedimento algumas vezes é referido como *padronização* da variável aleatória X, e Z é chamado uma **variável aleatória padronizada**. (Em cursos introdutórios de estatística, ele algumas vezes é chamado de *transformação-z* de X). É importante lembrar que o desvio-padrão, não a variância, aparece no denominador de (B.25). Como veremos, essa transformação é freqüentemente utilizada na inferência estatística.

Como um exemplo específico, suponha que $\text{E}(X) = 2$ e $\text{Var}(X) = 9$. Então, $Z = (X - 2)/3$ terá um valor esperado igual a zero e variância igual a um.

B.4 CARACTERÍSTICAS DAS DISTRIBUIÇÕES CONJUNTAS E CONDICIONAIS

Medidas de Associação: Covariância e Correlação

Embora a fdp conjunta de duas variáveis aleatórias descreva completamente a relação entre elas, é útil ter medidas resumidas de como, em média, duas variáveis aleatórias variam entre si. Como acontece com o valor esperado e a variância, isso é semelhante ao usar um único número para resumir alguma coisa de uma distribuição inteira, que, nesse caso, é uma distribuição conjunta de duas variáveis aleatórias.

Covariância

Seja $\mu_X = \text{E}(X)$, e $\mu_Y = \text{E}(Y)$, e considere a variável aleatória $(X - \mu_X)(Y - \mu_Y)$. Agora, se X e Y estiveram acima de suas respectivas médias, então, $(X - \mu_X)(Y - \mu_Y) > 0$. Isso também será verdadeiro se $X < \mu_X$ e $Y < \mu_Y$. Por outro lado, se $X > \mu_X$ e $Y < \mu_Y$, ou vice-versa, então, $(X - \mu_X)(Y - \mu_Y) < 0$. Como, então, esse produto poderá nos dar qualquer informação sobre a relação entre X e Y?

A **covariância** entre duas variáveis aleatórias X e Y, algumas vezes chamada a covariância populacional para enfatizar que ela se refere à relação entre duas variáveis descrevendo uma população, é definida como o valor esperado do produto $(X - \mu_X)(Y - \mu_Y)$:

$$\text{Cov}(X,Y) \equiv \text{E}[X - \mu_X)(Y - \mu_Y)], \tag{B.26}$$

que algumas vezes é representado por σ_{XY}. Se $\sigma_{XY} > 0$, então, em média, quando X estiver acima de sua média, Y também estará acima de sua média. Se $\sigma_{XY} < 0$, então, em média, quando X estiver acima de sua média, Y estará abaixo de sua média.

Algumas expressões úteis para calcular $\text{Cov}(X,Y)$ são as seguintes:

$$\begin{aligned}\text{Cov}(X,Y) &= \text{E}[X - \mu_X)(Y - \mu_Y)] = \text{E}[(X - \mu_X)Y] \\ &= \text{E}[X(Y - \mu_Y)] = \text{E}(XY) - \mu_X\mu_Y.\end{aligned} \tag{B.27}$$

Decorre de (B.27) que, se $\text{E}(X) = 0$ ou $\text{E}(Y) = 0$, então, $\text{Cov}(X,Y) = \text{E}(X,Y)$.

A covariância mede o grau de dependência *linear* entre duas variáveis aleatórias. Uma covariância positiva indica que duas variáveis aleatórias se movem na mesma direção, enquanto uma covariância negativa indica que elas se movem em direções opostas. Interpretar a *magnitude* de uma covariância pode ser um pouco difícil, como veremos brevemente.

Como a covariância é uma medida de como duas variáveis aleatórias estão relacionadas, é natural perguntar como a covariância está relacionada à noção de independência. Isso é dado pela seguinte propriedade.

PROPRIEDADE COV.1
Se X e Y forem independentes, então, $\text{Cov}(X,Y) = 0$.

Essa propriedade decorre da equação (B.27) e do fato de que $\text{E}(XY) = \text{E}(X)\text{E}(Y)$ quando X e Y são independentes. É importante lembrar que a inversa de COV.1 *não* é verdadeira: covariância zero entre X e Y não implica que X e Y sejam independentes. De fato, existem variáveis aleatórias X de tal forma que, se $Y = X^2$, $\text{Cov}(X,Y) = 0$. [Qualquer variável aleatória com $\text{E}(X) = 0$ e $\text{E}(X^3) = 0$ tem essa propriedade]. Se $Y = X^2$, então, X e Y serão claramente não-independentes: conhecendo X conheceremos Y. Parece bastante estranho que X e X^2 possam ter covariância zero, e isso revela um ponto fraco da covariância como uma medida geral de associação entre duas variáveis aleatórias. A covariância é útil em contextos em que as relações são pelo menos aproximadamente lineares.

A segunda mais importante propriedade da covariância envolve covariâncias entre funções lineares.

PROPRIEDADE COV.2
Para quaisquer constantes a_1, b_1 , a_2 e b_2,

$$\text{Cov}(a_1X + b_1, a_2Y + b_2) = a_1a_2\text{Cov}(X,Y). \qquad \textbf{(B.28)}$$

Uma implicação importante de COV.2 é que a covariância entre duas variáveis aleatórias pode ser alterada simplesmente pela multiplicação de uma ou de ambas as variáveis aleatórias por uma constante. Isso é importante em economia, pois variáveis monetárias, taxas de inflação etc. podem ser definidas com diferentes unidades de medida sem que seja alterado seu significado.

Finalmente, é útil saber que o valor absoluto da covariância entre quaisquer duas variáveis aleatórias está limitado pelo produto de seus desvios-padrão; isso é conhecido como a *desigualdade de Cauchy-Schwartz*.

PROPRIEDADE COV.3
$|\text{Cov}(X,Y)| \leq \text{dp}(X)\text{dp}(Y)$.

Coeficiente de Correlação

Suponha que queremos conhecer a relação entre o grau de educação e os rendimentos anuais da população que trabalha. Poderíamos chamar de X a educação e de Y os rendimentos, computando a seguir sua covariância. Entretanto, a resposta que obteremos dependerá de como escolheremos medir a educação e os rendimentos. A propriedade COV.2 implica que a covariância entre educação e rendimentos dependerá de se os rendimentos são medidos em dólares ou milhares de dólares, ou se a educação é medida em meses ou anos. É bastante claro que a maneira como mediremos essas variáveis não terá influência no quanto elas estão fortemente relacionadas. Mas a covariância entre elas efetivamente depende das unidades de medida.

O fato de a covariância depender das unidades de medida é uma deficiência que é compensada pelo **coeficiente de correlação** entre X e Y:

$$\text{Corr(X,Y)} \equiv \frac{\text{Cov}(X,Y)}{\text{dp}(X)\cdot\text{dp}(Y)} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}; \tag{B.29}$$

o coeficiente de correlação entre X e Y algumas vezes é representado por ρ_{XY} (e é algumas vezes chamado de correlação populacional).

Como σ_X e σ_Y são positivos, $\text{Cov}(X,Y)$ e $\text{Corr}(X,Y)$ sempre têm o mesmo sinal, e $\text{Corr}(X,Y) = 0$ se, e somente se, $\text{Cov}(X,Y) = 0$. Algumas das propriedades da covariância são transferidas para a correlação. Se X e Y forem independentes, então, $\text{Corr}(X,Y) = 0$, mas correlação zero não implica independência. (Como a covariância, o coeficiente de correlação também é uma medida de dependência linear). Porém, a magnitude do coeficiente de correlação é mais fácil de interpretar do que o tamanho da covariância, devido à seguinte propriedade.

PROPRIEDADE CORR.1
$-1 \leq \text{Corr}(X,Y) \leq 1$.

Se $\text{Corr}(X,Y) = 0$ ou, equivalentemente, $\text{Cov}(X,Y) = 0$, não haverá relação linear entre X e Y, e X e Y são chamadas de variáveis *não-correlacionadas*; caso contrário, X e Y serão correlacionadas. $\text{Corr}(X,Y) = 1$ indica uma relação linear positiva perfeita, o que significa que podemos escrever $Y = a + bX$ para alguma constante a e alguma constante $b > 0$. $\text{Corr}(X,Y) = -1$ indica uma relação linear negativa perfeita, de forma que $Y = a + bX$ para alguma constante $b < 0$. Os casos extremos de correlação positiva ou negativa igual à unidade raramente ocorre. Valores de RHO_{XY} próximos de 1 ou -1 indicam fortes relações lineares.

Como mencionado antes, a correlação entre X e Y não varia em relação às unidades de medida de X ou Y. Isso é especificado de forma mais geral como segue.

PROPRIEDADE CORR.2
Para as constantes a_1, b_1, a_2 e b_2, com $a_1a_2 > 0$,

$$\text{Corr}(a_1\text{X} + b_1, a_2Y + b_2) = \text{Corr}(X,Y).$$

Se $a_1a_2 < 0$, então,

$$\text{Corr}(a_1\text{X} + b_1, a_2Y + b_2) = -\text{Corr}(X,Y).$$

Como um exemplo, suponha que a correlação entre rendimentos e educação da população que trabalha seja 0,15. Essa medida não depende de se os rendimentos estão medidos em dólares, milhares de dólares, ou qualquer outra unidade; ela também não depende de se a educação está medida em anos, trimestres, meses etc.

Variância da Soma de Variáveis Aleatórias

Agora que já definimos covariância e correlação, podemos completar nossa relação das principais propriedades da variância.

PROPRIEDADE VAR.3
Para as constante a e b,

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X,Y).$$

Segue imediatamente que, se X e Y forem não-correlacionadas — de forma que $\text{Cov}(X,Y) = 0$ —, então,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \qquad \textbf{(B.30)}$$

e

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y). \qquad \textbf{(B.31)}$$

Neste último caso, observe como a variância da diferença é a *soma*, não a diferença, das variâncias.

Como um exemplo de (B.30), seja X os lucros ganhos por um restaurante durante uma noite de sexta-feira e Y os lucros ganhos na noite do sábado seguinte. Então, $Z = X + Y$ será os lucros das duas noites. Suponha que X e Y tenham, cada uma, um valor esperado de 300 dólares e um desvio-padrão de 15 dólares (de forma que a variância será 255). O lucro esperado das duas noites será $\text{E}(Z) + \text{E}(Y) = 2\cdot(300) = 600$ dólares. Se X e Y forem independentes e, portanto, não-correlacionadas, a variância do lucro total será a soma das variâncias: $\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y) = 2\cdot(225) = 450$. Portanto, o desvio-padrão do lucro total será $\sqrt{450}$ ou aproximadamente 21,21 dólares.

As expressões (B.30) e (B.31) estendem-se para mais de duas variáveis aleatórias. Para especificar essa extensão, precisamos de uma definição. As variáveis aleatórias X e Y serão **variáveis aleatórias não-correlacionadas duas a duas** se cada variável no conjunto for não-correlacionada com cada outra variável do conjunto. Isto é, $\text{Cov}(X_i,X_j) = 0$ para todo $i \neq j$.

PROPRIEDADE VAR.4
Se $\{X_1, ..., X_n)$ forem variáveis aleatórias não-correlacionadas duas a duas e $\{a_i: i = 1, ..., n\}$ forem constantes, então,

$$\text{Var}(a_1X_1 + ... + a_nX_n) = a_1^2\text{Var}(X_1) + ... + a_n^2\text{Var}(X_n).$$

Em notação de somatórios, podemos escrever

$$\text{Var}\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2\text{Var}(X_i). \qquad \textbf{(B.32)}$$

Um caso especial da Propriedade VAR.4 ocorre quando consideramos $a_i = 1$ para todos os i. Então, para variáveis aleatórias não-correlacionadas duas a duas, a variância da soma será a soma das variâncias:

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i). \qquad \textbf{(B.33)}$$

Como variáveis aleatórias independentes são não-correlacionadas (veja a propriedade COV.1), a variância de uma soma de variáveis aleatórias independentes é a soma das variâncias.

Se as variáveis X_i não forem não-correlacionadas duas a duas, então, a expressão $\text{Var}(\sum_{i=1}^{n} a_i X_i)$ será muito mais complicada; precisaremos adicionar no lado direito de (B.32) os termos $2a_i a_j \text{Cov}(x_i, x_j)$ para todo $i > j$.

Podemos usar (B.33) para derivarmos a variância de uma variável aleatória binomial. Definimos $X \sim \text{Binomial}(n,\theta)$ e escrevemos $X = Y_1, + ... + Y_n$, onde Y_i são variáveis aleatórias independentes de Bernoulli(θ). Então, de (B.33), $\text{Var}(X) = \text{Var}(Y_1) + ... + \text{Var}(Y_n) = n\theta(1 - \theta)$.

No exemplo das reservas da companhia aérea com $n = 120$ e $\theta = 0{,}85$, a variância do número de passageiros que comparecem para embarque seria $120(0{,}85)(0{,}15) = 15{,}3$, e, assim, o desvio-padrão seria aproximadamente 3,9.

Esperança Condicional

A covariância e a correlação medem a relação linear entre duas variáveis aleatórias e as tratam simetricamente. Muitas vezes, em ciências sociais, gostaríamos de explicar uma variável, chamada Y, em termos de outra variável, digamos X. Além disso, se Y for relacionada com X de uma maneira não linear, gostaríamos de ser informados sobre isso. Chamemos Y de variável explicada e X de variável explicativa. Por exemplo, Y poderia ser o salário por hora e X poderia ser o número de anos de educação formal.

Já introduzimos a noção de função de densidade de probabilidade condicional de Y, dado X. Assim, poderíamos querer ver como a distribuição dos salários é alterada pelo nível de educação. Porém, em geral, queremos ter uma maneira simples de resumir essa distribuição. Um único número não será suficiente, visto que a distribuição de Y, dado $X = x$, geralmente depende do valor de x. No entanto, podemos resumir a relação entre Y e X verificando a **esperança condicional** de Y, dado X, algumas vezes chamada *média condicional*. A idéia é a seguinte: suponha que saibamos que X assumiu um valor particular, digamos x. Então, poderemos calcular o valor esperado de Y em decorrência de conhecermos esse resultado de X. Representamos esse valor esperado por $\text{E}(Y|X = x)$, ou algumas vezes $\text{E}(Y|x)$ como forma abreviada. De forma geral, quando x muda, $\text{E}(Y|x)$ também muda.

Quando Y for uma variável aleatória discreta assumindo valores $\{y_1, ..., y_n\}$, então,

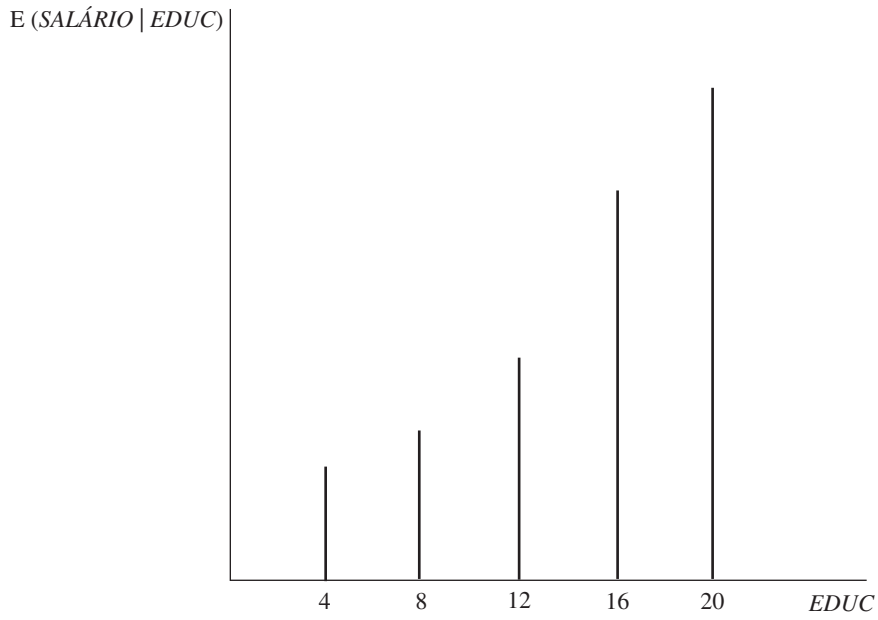
$$\text{E}(Y|x) = \sum_{j=1}^{m} y_j f_{Y|X}(y_j|x).$$

Quando Y for contínua, $\text{E}(Y|x)$ será definida pela integração de $yf_{Y|X}(y|x)$ sobre todos os valores possíveis de y. Assim como no caso da esperança incondicional, a esperança condicional é uma média ponderada de possíveis valores de Y, mas agora os pesos refletem o fato de que X assumiu um valor específico. Assim, $\text{E}(Y|x)$ é apenas alguma função de x, que nos diz como o valor esperado de Y varia com x.

Como um exemplo, seja (X,Y) a população de todas as pessoas que trabalham, na qual X é anos de educação, e Y é o salário por hora. Então, $\text{E}(Y|X = 12)$ será o salário médio por hora de todas as pessoas da população com 12 anos de educação (em termos gerais, correspondente à educação de ensino médio). $\text{E}(Y|X = 16)$ será o salário médio por hora de todas as pessoas com 16 anos de educação. O gráfico de valores esperados com vários níveis de educação fornece informações importantes sobre como os salários e a educação estão relacionados. Veja a Figura B.5, para uma ilustração.

Figura B.5

O valor esperado do salário por hora considerando vários níveis de educação.



Em princípio, o valor esperado do salário por hora pode ser encontrado a cada nível de educação, e essas esperanças podem ser resumidas em uma tabela. Como a educação pode variar amplamente — e pode até mesmo ser medida em frações de um ano —, essa é uma maneira excessivamente trabalhosa de se mostrar a relação entre o salário médio e o grau de educação. Em econometria, geralmente especificamos funções simples que capturam essa relação. Como um exemplo, suponha que o valor esperado de *SALÁRIO*, dado *EDUC*, seja a função linear

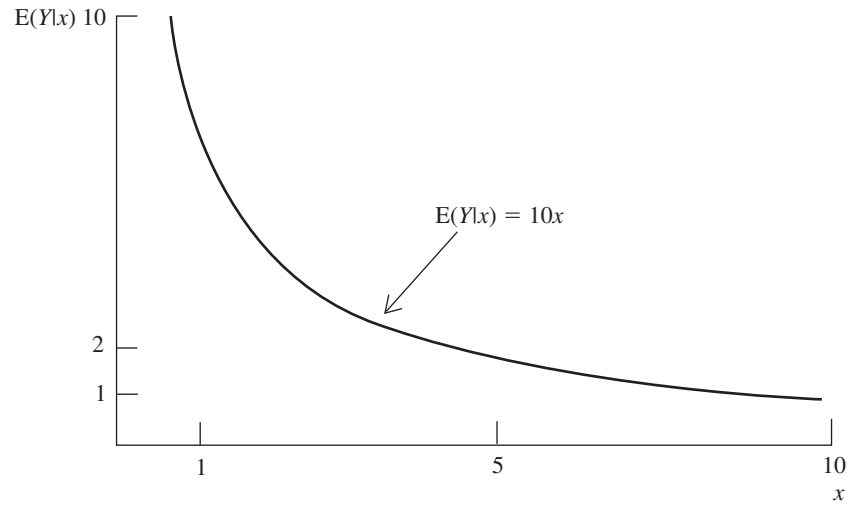
$$\mathrm{E}(SALÁRIO|EDUC) = 1{,}05 + 0{,}45\ EDUC.$$

Se essa relação for válida na população das pessoas que trabalham, o salário médio das pessoas com 8 anos de educação será 1,05 + 0,45(8) = 4,65, ou 4,65 dólares. O salário médio das pessoas com 16 anos de educação será 8,25 dólares. O coeficiente de *EDUC* implica que cada ano de educação aumenta o salário por hora esperado em 0,45, ou 45 centavos de dólar.

As esperanças condicionais também podem ser funções não-lineares. Por exemplo, suponha que $\mathrm{E}(Y|x) = 10/x$, onde X é uma variável aleatória que sempre será maior que zero. Essa função está traçada na Figura B.6. Isso poderia representar uma função de demanda, na qual Y seria a quantidade demandada e X seria o preço. Se Y e X forem relacionadas nesta forma, uma análise de associação linear, tal como uma análise de correlação, seria incompleta.

Figura B.6

Gráfico de $E(Y|x) = 10/x$.



Propriedades da Esperança Condicional

Várias propriedades básicas das esperanças condicionais são úteis para derivações em análise econométrica.

PROPRIEDADE EC.1

$E[c(X)|X] = c(X)$, para qualquer função $c(X)$.

Essa primeira propriedade significa que funções de X comportam-se como constantes quando calculamos a esperança condicional de X. Por exemplo, $E(X^2|X) = X^2$. Intuitivamente, isso simplesmente significa que, se conhecermos X, também conheceremos X^2.

PROPRIEDADE EC.2

Para as funções $a(X)$ e $b(X)$,

$$E[a(X)Y + b(X)|X] = a(X)E(Y|X) + b(X).$$

Por exemplo, podemos calcular com facilidade a esperança condicional de uma função tal como $XY + 2X^2$: $E(XY + 2X^2|X)$ $XE(Y|X) + 2X^2$.

A próxima propriedade interliga as noções de independência e esperanças condicionais.

PROPRIEDADE EC.3

Se X e Y forem independentes, então, $E(Y|X) = E(Y)$.

Essa propriedade significa que, se X e Y forem independentes, então, o valor esperado de Y, dado X, não dependerá de X, caso em que $E(Y|X)$ sempre será igual ao valor esperado (incondicional) de Y. No exemplo do salário e educação, se salário fosse independente de educação, então, os salários médios

das pessoas com educação de ensino médio e com cursos superiores seriam os mesmos. Como quase certamente esse resultado seria falso, não podemos assumir que salário e educação sejam independentes.

Um caso especial da propriedade EC.3 é o seguinte: se U e X forem independentes e $E(U) = 0$, então, $E(U|X) = 0$.

Também existem propriedades da esperança condicional que têm a ver com o fato de $E(Y|X)$ ser uma função de X, digamos $E(Y|X) = \mu(X)$. Como X é uma variável aleatória, $\mu(X)$ também será uma variável aleatória. Além disso, $\mu(X)$ tem uma distribuição de probabilidade e, portanto, um valor esperado. De forma geral, o valor esperado de $\mu(X)$ pode ser muito difícil de ser calculado de forma direta. A **lei das expectativas iteradas** diz que o valor esperado de $\mu(X)$ é simplesmente igual ao valor esperado de Y. Escrevemos isso da seguinte maneira.

PROPRIEDADE EC.4

$E[E(Y|X)] = E(Y)$.

Essa propriedade é de difícil compreensão à primeira vista. Ela significa que, se primeiro obtivermos $E(Y|X)$ como uma função de X e considerarmos seu valor esperado (em relação à distribuição de X, é claro), então, acabaremos obtendo $E(Y)$. Isso não é tão óbvio, mas pode ser derivado utilizando a definição dos valores esperados.

Suponha que $Y = SALÁRIO$ e $X = EDUC$, onde *SALÁRIO* está medido em horas e *EDUC* em anos. Suponha que o valor esperado de *SALÁRIO*, dado *EDUC*, seja $E(SALÁRIO|EDUC) = 4 + 0{,}60\ EDUC$. Além disso, $E(EDUC) = 11{,}5$. Então, a lei das expectativas iteradas sugere que $E(SALÁRIO) = E(4 + 0{,}60\ EDUC) = 4 + 0{,}60\ E(EDUC) = 4 + 0{,}60(11{,}5) = 10{,}90$, ou 10,90 dólares por hora.

A próxima propriedade especifica uma versão mais geral da lei das expectativas iteradas.

PROPRIEDADE EC.4′

$E(Y|X) = E[E(Y|X,Z)|X]$.

Em outras palavras, podemos encontrar $E(Y|X)$ em duas etapas. Primeiro, encontramos $E(Y|X,Z)$ para qualquer outra variável aleatória Z. Em seguida, encontramos o valor esperado de $E(Y|X,Z)$, condicional em X.

PROPRIEDADE EC.5

Se $E(Y|X) = E(Y)$, então, $Cov(X,Y) = 0$ [como também $Corr(X,Y) = 0$]. De fato, *qualquer* função de X é não-correlacionada com Y.

Essa propriedade significa que, se o conhecimento de X não altera o valor esperado de Y, então, X e Y *devem* ser não-correlacionadas, o que implica que, se X e Y forem correlacionadas, então, $E(Y|X)$ deve depender de X. A inversa da propriedade EC.5 não é verdadeira: se X e Y forem não-correlacionadas, $E(Y|X)$ *poderá* ainda depender de X. Por exemplo, suponha que $Y = X^2$. Então, $E(Y|X) = X^2$, que claramente é uma função de X. Porém, como mencionado em nossa discussão sobre covariância e correlação, é possível que X e X^2 sejam não-correlacionadas. A esperança condicional captura a relação não linear entre X e Y que uma análise de correlação deixaria passar despercebida.

As propriedades EC.4 e EC.5 têm duas implicações importantes: se U e X forem variáveis aleatórias, de forma que $E(U|X) = 0$, então, $E(U) = 0$, e U e X serão não-correlacionadas.

PROPRIEDADE EC.6
Se $E(Y^2) < \infty$ e $E[g(X)^2] < \infty$ para alguma função g, então, $E\{[Y - \mu(X)]^2|X\} \leq E\{[Y - g(X)]^2|X\}$ e $E\{[Y - \mu(X)]^2\} \leq E\{[Y - g(X)]^2\}$.

A propriedade EC.6 é muito útil em contextos de previsão ou de projeções. A primeira desigualdade diz que, se medirmos a inexatidão da previsão como o erro quadrático de previsão esperado, condicional em X, então, a média condicional será melhor que qualquer outra função de X para prever Y. A média condicional também minimiza o erro quadrático de previsão esperado incondicional.

Variância Condicional

Dadas as variáveis aleatórias X e Y, a variância de Y, condicional em $X = x$, será simplesmente a variância associada à distribuição condicional de Y, dado $X = x$: $E\{[Y - E(Y|x)]^2|x\}$. A fórmula

$$Var(Y|X = x) = E(Y^2|x) - [E(Y|x)]^2$$

é freqüentemente útil para os cálculos. Somente ocasionalmente teremos que calcular uma variância condicional. Entretanto, teremos que fazer hipóteses a respeito e manipular as variâncias condicionais para certos tópicos na análise de regressão.

Como um exemplo, defina $Y = POUPANÇA$ e $X = RENDA$ (ambas medidas em termos anuais, para a população de todas as famílias). Suponha que $Var(POUPANÇA|RENDA) = 400 + 0{,}25\ RENDA$. Isso diz que, conforme aumente a renda, a variância dos níveis de poupança também aumenta. É importante verificar que a relação entre as variâncias de *POUPANÇA* e *RENDA* é totalmente separada da relação entre os *valores esperados* de *POUPANÇA* e *RENDA*.

Estabelecemos, portanto, uma propriedade importante da variância condicional.

PROPRIEDADE VC.1
Se X e Y forem independentes, então, $Var(Y|X) = Var(Y)$.

Essa propriedade é bastante clara, pois a distribuição de Y, dado X, não depende de X, e $Var(Y|X)$ é apenas uma característica dessa distribuição.

B.5 A DISTRIBUIÇÃO NORMAL E OUTRAS DISTRIBUIÇÕES A ELA RELACIONADAS

A Distribuição Normal

A distribuição normal e as derivadas dela são as distribuições mais amplamente usadas em estatística e econometria. Assumir que variáveis aleatórias definidas para populações são normalmente distribuídas simplifica os cálculos de probabilidade. Além disso, nos valeremos pesadamente da distribuição normal e de outras a ela relacionadas para conduzir inferência em estatística e econometria — mesmo quando a população básica não for necessariamente normal. Precisamos adiar os detalhes, mas tenha certeza de que essas distribuições irão surgir muitas vezes ao longo deste livro.

Uma variável aleatória normal é uma variável aleatória contínua que pode assumir qualquer valor. Sua função de densidade de probabilidade tem a forma familiar de um sino traçada na Figura B.7.

Matematicamente, a fdp de X pode ser escrita como

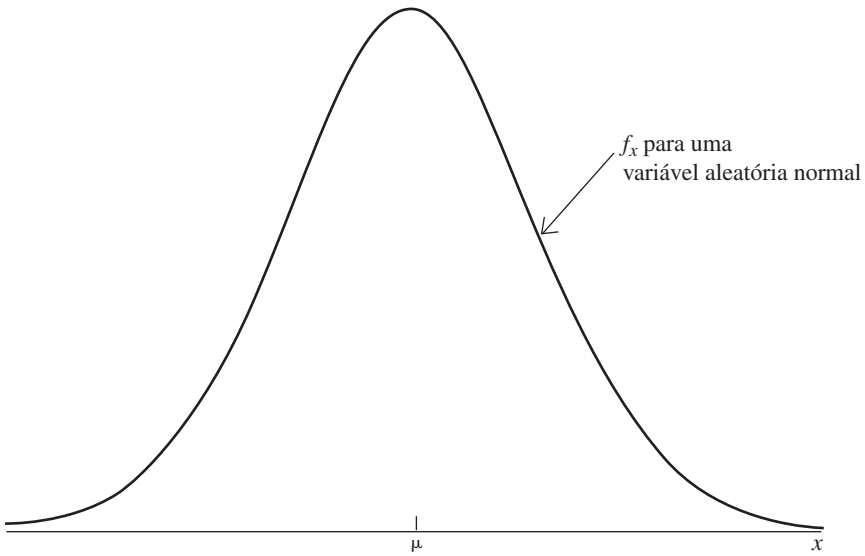
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(x - \mu)^2/2\sigma^2], \; -\infty < x < \infty, \qquad \textbf{(B.34)}$$

onde $\mu = \mathrm{E}(X)$ e $\sigma^2 = \mathrm{Var}(X)$. Dizemos que X tem uma **distribuição normal** com valor esperado μ e variância σ^2, escrita como $X \sim \mathrm{Normal}(\mu,\sigma^2)$. Como a distribuição normal é simétrica em relação a μ, μ também é a mediana de X. A distribuição normal é algumas vezes chamada de *distribuição de Gauss* em homenagem ao famoso estatístico C. F. Gauss.

Algumas variáveis aleatórias parecem seguir em linhas gerais uma distribuição normal. As alturas e pesos do ser humano, resultados de provas e índices de desemprego municipais possuem fdps com aproximadamente a forma na Figura B.7. Outras distribuições, como as da renda, não parecem seguir a função de probabilidade normal. Na maioria dos países, a renda não é simetricamente distribuída em torno de qualquer valor; a distribuição é inclinada em direção à extremidade superior. Em alguns casos, uma variável pode ser transformada para atingir a normalidade. Uma transformação popular é o log natural, que faz sentido para variáveis aleatórias positivas. Se X for uma variável aleatória positiva, tal como a renda, e $Y = \log(X)$ tiver uma distribuição normal, então, dizemos que X tem uma *distribuição lognormal*. É possível constatar que a distribuição lognormal se ajusta bastante bem à distribuição de renda de muitos países. Outras variáveis, como preços de mercadorias, parecem ser bem descritas quando utilizada a distribuição lognormal.

Figura B.7

A forma geral de uma função de densidade de probabilidade normal.



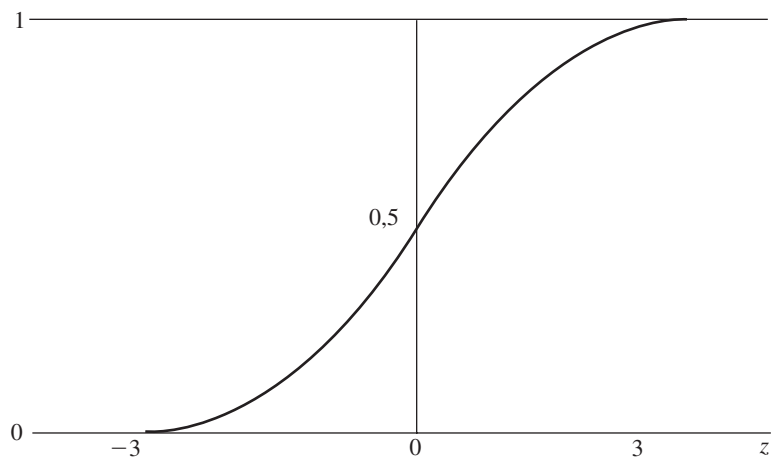
A Distribuição Normal Padrão

Um caso especial da distribuição normal ocorre quando a média for zero e a variância (e, portanto, o desvio-padrão) for a unidade. Se uma variável aleatória Z tiver uma distribuição Normal(0,1), dizemos que ela tem um **distribuição normal padrão**. A fdp de uma variável aleatória normal padrão é representada por PHI(z); de (B.34), com $\mu = 0$ e $\sigma^2 = 1$, ela é dada por

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2), \; -\infty < z < \infty. \tag{B.35}$$

A função de distribuição cumulativa normal padrão é representada por $\Phi(z)$ e é obtida como a área sob ϕ, à esquerda de z; veja a Figura B.8. Lembre-se de que $\Phi(z) = P(Z \leq z)$; como Z é contínua, $\Phi(z) = P(Z < z)$ também é contínua.

Figura B.8
A função de distribuição cumulativa normal padrão.



Não existe fórmula simples que possa ser usada para obter os valores de $\Phi(z)$ [pois $\Phi(z)$ é a integral da função em (B.35), e essa integral não tem forma única]. No entanto, os valores de $\Phi(z)$ são facilmente tabulados; eles estão dados para z entre $-3{,}1$ e 3,1 na Tabela G.1 no Apêndice G. Para $z \leq -3{,}1$, $\Phi(z)$ será menor que 0,001, e para $z \geq 3{,}1$, $\Phi(z)$ será maior que 0,999. A maioria dos programas estatísticos e econométricos inclui comandos simples para computar valores da fdc normal padrão, e assim podemos freqüentemente evitar totalmente as tabelas impressas para obter as probabilidades para qualquer valor de z.

Usando fatos básicos da probabilidade — e, em particular, as propriedades (B.7) e (B.8) em relação às fdcs —, podemos usar a fdc normal padrão para calcular a probabilidade de qualquer evento envolvendo uma variável aleatória normal padrão. As fórmulas mais importantes são

$$P(Z > z) = 1 - \Phi(z), \tag{B.36}$$

$$P(Z < -z) = P(Z > z) \tag{B.37}$$

e

$$P(a \leq Z \leq b) = \Phi(b) - \Phi(a). \tag{B.38}$$

Como Z é uma variável aleatória contínua, todas as três fórmulas são válidas, sejam ou não restritas as desigualdades. Citamos alguns exemplos: $P(Z > 0{,}44) = 1 - 0{,}67 = 0{,}33$, $P(Z < -0{,}92) = P(Z > 0{,}92) = 1 - 0{,}821 = 0{,}179$, e $P(-1 < Z \leq 0{,}5) = 0{,}692 - 0{,}159 = 0{,}533$.

Outra expressão útil é que, para qualquer $c > 0$,

$$\begin{aligned} P(|Z| > c) &= P(Z > c) + P(Z < -c) \\ &= 2 \cdot P(Z > c) = 2[1 - \Phi(c)]. \end{aligned} \tag{B.39}$$

Assim, a probabilidade de que o valor absoluto de Z seja maior que alguma constante c positiva será simplesmente duas vezes a probabilidade $P(Z > c)$; isso reflete a simetria da distribuição normal padrão.

Na maioria das aplicações, iniciamos com uma variável aleatória normalmente distribuída, $X \sim \text{Normal}(\mu,\sigma^2)$, onde μ é diferente de zero e $\sigma^2 \neq 1$. Qualquer variável aleatória normal pode ser transformada em uma normal padrão usando a seguinte propriedade.

PROPRIEDADE NORMAL.1

Se $X \sim \text{Normal}(\mu,\sigma^2)$, então, $(X - \mu)/\sigma \sim \text{Normal}(0,1)$.

A propriedade Normal.1 mostra como transformar qualquer variável aleatória normal em uma normal padrão. Assim, suponha que $X \sim \text{Normal}(3,4)$, e que gostaríamos de calcular $P(X \leq 1)$. As etapas sempre envolvem a normalização de X para uma normal padrão:

$$\begin{aligned} P(X \leq 1) &= P(X - 3 \leq 1 - 3) = P\left(\frac{X - 3}{2} \leq -1\right) \\ &= P(Z \leq -1) = \Phi(-1) = 0{,}159. \end{aligned}$$

EXEMPLO B.6

(Probabilidades para uma Variável Aleatória Normal)

Primeiro, vamos calcular $P(2 < X \leq 6)$ quando $X \sim \text{Normal}(4,9)$ (usar $<$ ou $\leq$ é irrelevante, pois X é uma variável aleatória contínua). Agora

$$\begin{aligned} P(2 < X \leq 6) &= P\left(\frac{2 - 4}{3} < \frac{X - 4}{3} \leq \frac{6 - 4}{3}\right) = P(-2/3 < Z \leq 2/3) \\ &= \Phi(0{,}67) - \Phi(-0{,}67) = 0{,}749 - 0{,}251 = 0{,}498. \end{aligned}$$

Agora, vamos calcular $P(|X| > 2)$:

$$\begin{aligned} P(|X| > 2) &= P(X > 2) + P(X < -2) \\ &= P[(X - 4)/3 > (2 - 4)/3] + P[(X - 4)/3 < (-2 - 4)/3] \\ &= 1 - \Phi(-2/3) + \Phi(-2) \\ &= 1 - 0{,}251 + 0{,}023 = 0{,}772. \end{aligned}$$

Propriedades Adicionais da Distribuição Normal

Terminamos esta subseção reunindo vários outros fatos sobre as distribuições normais que usaremos mais tarde.

PROPRIEDADE NORMAL.2
Se $X \sim \text{Normal}(\mu,\sigma^2)$, então, $aX + b \sim \text{Normal}(a\mu + b, a^2\mu^2)$.

Assim, se $X \sim \text{Normal}(1,9)$, então, $Y = 2X + 3$ será distribuída como normal com média $2\text{E}(X) + 3 = 5$ e variância $2^2 \cdot 9 = 36$; $\text{dp}(Y) = 2\text{dp}(X) = 2 \cdot 3 = 6$.

Anteriormente, explicamos como, em geral, correlação zero e independência não são a mesma coisa. No caso de variáveis aleatórias normalmente distribuídas, é possível constatar que a correlação zero é suficiente para a independência.

PROPRIEDADE NORMAL.3
Se X e Y forem conjunta e normalmente distribuídas, então, elas serão independentes se, e somente se, $\text{Cov}(X,Y) = 0$.

PROPRIEDADE NORMAL.4
Qualquer combinação linear de variáveis aleatórias independentes e identicamente normalmente distribuídas tem uma distribuição normal.

Por exemplo, sejam X_i, para $i = 1$, 2 e 3, variáveis aleatórias independentes distribuídas como $\text{Normal}(\mu,\sigma^2)$. Defina $W = X_1 + 2X_2 - 3X_3$. Então, W será normalmente distribuída; precisamos simplesmente encontrar sua média e sua variância. Agora,

$$\text{E}(W) = \text{E}(X_1) + 2\text{E}(X_2) - 3\text{E}(X_3) = \mu + 2\mu - 3\mu = 0.$$

Além disso,

$$\text{Var}(W) = \text{Var}(X_1) + 4\text{Var}(X_2) + 9\text{Var}(X_3) = 14\sigma^2.$$

A propriedade Normal.4 também conclui que a média das variáveis aleatórias independentes normalmente distribuídas tem uma distribuição normal. Se $Y_1, Y_2, \ldots, Y_n$ forem variáveis aleatórias independentes e cada uma for distribuída como $\text{Normal}(\mu,\sigma^2)$, então,

$$\overline{Y} \sim \text{Normal}(\mu,\sigma^2/n). \qquad \textbf{(B.40)}$$

Esse resultado é crítico para a inferência com respeito à média em uma população normal.

A Distribuição Qui-Quadrado

A distribuição qui-quadrado é obtida diretamente das variáveis aleatórias independentes normais padrões. Sejam Z_i, $i = 1,2,\ldots,n$ variáveis aleatórias independentes, cada uma distribuída como normal padrão. Defina uma nova variável aleatória como a soma dos quadrados de Z_i:

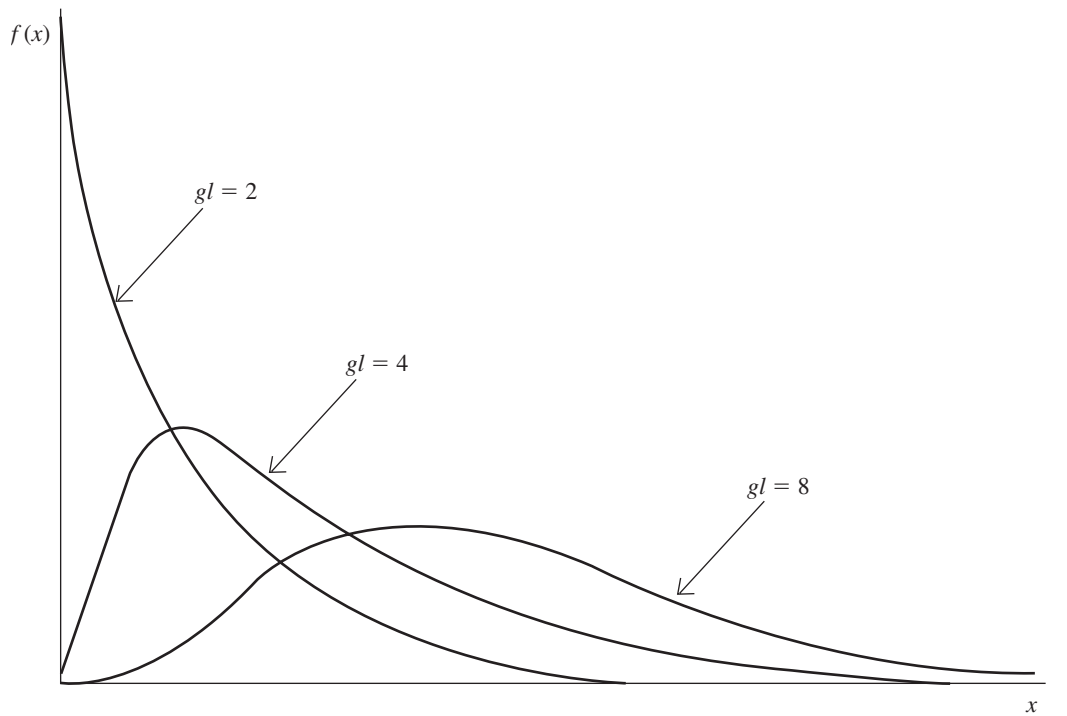
$$\text{X} = \sum_{i=1}^{n} Z_i^2. \qquad \textbf{(B.41)}$$

Então, X terá o que é conhecido como **distribuição qui-quadrado** com n **graus de liberdade** (ou gl abreviadamente). Escrevemos isso como $X \sim \chi^2_n$. Os gl em uma distribuição qui-quadrado correspondem ao número de termos na soma em (B.41). O conceito de graus de liberdade desempenhará um papel importante em nossas análises estatísticas e econométricas.

A fdp de uma distribuição qui-quadrado com diversos graus de liberdade é mostrada na Figura B.9; não precisaremos da fórmula dessa fdp e, portanto, não a reproduzimos aqui. Pela equação (B.41), fica claro que uma variável aleatória qui-quadrada é sempre não-negativa, e que, diferentemente da distribuição normal, a distribuição qui-quadrado não é simétrica em torno de qualquer ponto. É possível mostrar que, se $X \sim \chi^2_n$, o valor esperado de X será n [o número de termos em (B.41)], e a variância de X será $2n$.

Figura B.9

A distribuição qui-quadrado com vários graus de liberdade.



A Distribuição *t*

A distribuição t é o burro de carga na estatística clássica e nas análises de regressão múltipla. Obtemos uma distribuição t a partir de uma variável aleatória normal padrão e de uma variável aleatória qui-quadrada.

Suponhamos que Z tenha uma distribuição normal padrão e que X tenha uma distribuição qui-quadrado com n graus de liberdade. Adicionalmente, suponhamos que Z e X sejam independentes. Então, a variável aleatória

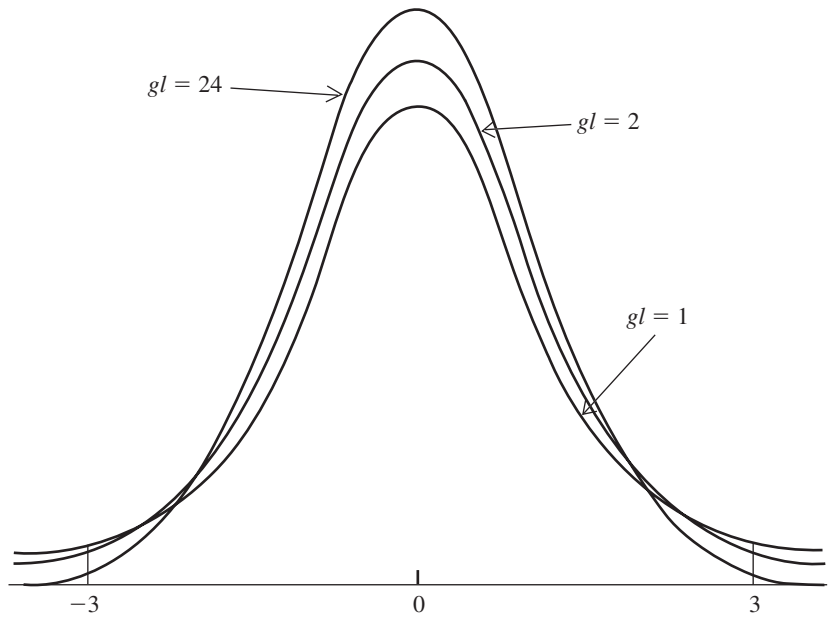
$$T = \frac{Z}{\sqrt{X/n}} \tag{B.42}$$

terá uma **distribuição *t*** com n graus de liberdade. Vamos representar isso como $T \sim t_n$. A distribuição t obtém seus graus de liberdade da variável aleatória qui-quadrada no denominador de (B.42).

A fdp da distribuição t tem uma forma semelhante à da distribuição normal padrão, exceto pelo fato de que ela é mais espalhada e, portanto, tem mais área nos extremos. O valor esperado de uma variável aleatória com distribuição t é zero (no sentido exato, o valor esperado somente existirá para $n > 1$), e a variância será $n/(n-2)$ para $n > 2$. (Não existe variância de $n \leq 2$ devido à distribuição ser tão espalhada.) A fdp da distribuição t está traçada na Figura B.10 para vários graus de liberdade. Conforme os graus de liberdade vão ficando maiores, a distribuição t se aproxima da distribuição normal padrão.

Figura B.10

A distribuição t com vários graus de liberdade.



A Distribuição *F*

Outra distribuição importante na estatística e na econometria é a distribuição F. Em particular, a distribuição F será usada para testar hipóteses no contexto de análise de regressão múltipla.

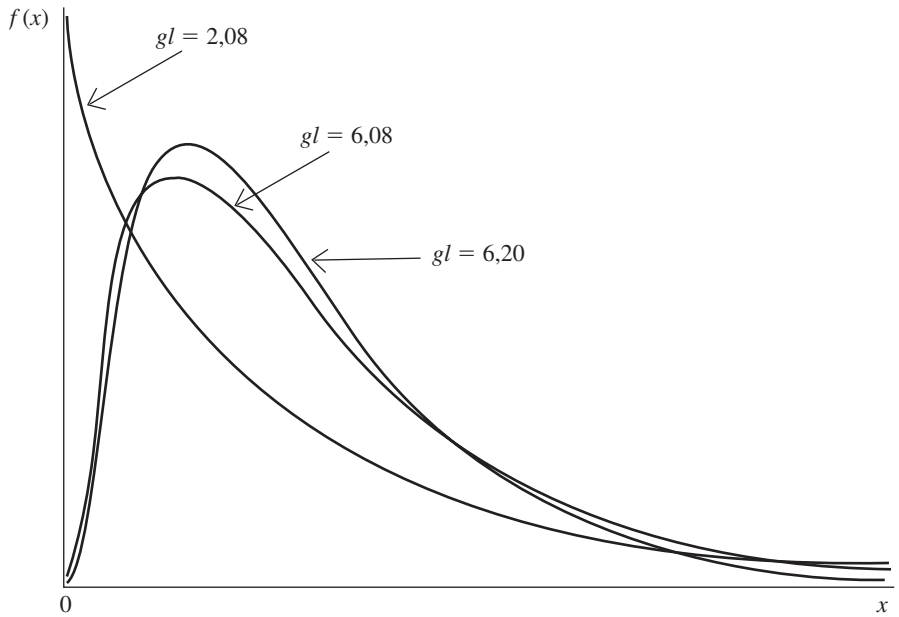
Para definir uma variável aleatória F, sejam $X_1 \sim \chi^2_{k_1}$ e $X_2 \sim \chi^2_{k_2}$ e X_1 e X_2 sejam independentes. Então, a variável aleatória

$$F = \frac{(X_1/k_1)}{(X_2/k_2)} \tag{B.43}$$

terá uma **distribuição *F*** com (k_1, k_2) graus de liberdade. Representamos isso como $F \sim F_{k_1,k_2}$. A fdp da distribuição *F* com diferentes graus de liberdade é mostrada na Figura B.11.

Figura B.11

A distribuição F_{k_1,k_2} para vários graus de liberdade, k_1 e k_2.



A ordem dos graus de liberdade em $F_{k1,k2}$ é crítica. O inteiro k_1 é chamado de *numerador dos graus de liberdade*, por ele estar associado à variável qui-quadrada no numerador. Da mesma forma, o inteiro k_2 é chamado de *denominador dos graus de liberdade*, por ele estar associado à variável qui-quadrada no denominador. Isso pode ser um pouco complicado, pois (B.43) também pode ser escrita como $(X_1k_2)/(X_2k_1)$, de forma que k_1 aparece no denominador. Apenas lembre-se de que o numerador *gl* é o inteiro associado à variável qui-quadrada no numerador de (B.43); de forma semelhante, uma associação é válida para o denominador *gl*.

RESUMO

Neste apêndice, revisamos os conceitos de probabilidade que são necessários em econometria. A maioria desses conceitos deve ser familiar a você de seus cursos introdutórios de probabilidade e estatística. Alguns dos tópicos mais avançados, como as características das esperanças condicionais, não precisam ser dominados agora — haverá tempo para tal quando esses conceitos surgirem, no contexto da análise de regressão na Parte 1.

Em um curso introdutório de estatística, o foco está no cálculo de médias, variâncias, covariâncias, e assim por diante, para distribuições particulares. Na Parte 1, não precisaremos de tais cálculos:

na maioria das vezes, recorremos às *propriedades* das esperanças, das variâncias etc., que explicamos neste apêndice.

PROBLEMAS

B.1 Suponha que um aluno do ensino médio está se preparando para prestar o exame vestibular. Explique por que a nota do vestibular dele é adequadamente vista como uma variável aleatória.

B.2 Defina X como uma variável aleatória distribuída como Normal(5,4). Encontre as probabilidades dos seguintes eventos

(i) $P(X \leq 6)$

(ii) $P(X > 4)$

(iii) $P(|X - 5| > 1)$

B.3 Muito se fala sobre o fato de que certos fundos mútuos têm desempenho superior ao do mercado ano após ano (isto é, o retorno por manter quotas nos fundos mútuos é maior que o retorno de possuir um portfólio como o da S&P 500). Em termos concretos, considere um período de dez anos e que a população de fundos mútuos seja os 4.170 reportados no *The Wall Street Journal* de 1º de janeiro de 1995. Ao dizermos que o desempenho relativo ao mercado é aleatório, queremos dizer que cada fundo tem uma possibilidade 50-50 de ter desempenho superior ao do mercado em qualquer ano e que o desempenho é independente de ano para ano.

(i) Se o desempenho relativo ao mercado for realmente aleatório, qual será a probabilidade de que qualquer fundo particular tenha um desempenho superior ao do mercado em todos os dez anos?

(ii) Encontre a probabilidade de que pelo menos um fundo dos 4.170 tenha um desempenho superior ao do mercado em todos os dez anos. Qual sua conclusão sobre sua resposta?

(iii) Se você possuir um programa estatístico que calcule probabilidades binomiais, encontre a probabilidade de que pelo menos cinco fundos tenham desempenho superior ao do mercado em todos os dez anos.

B.4 Para um município selecionado aleatoriamente nos Estados Unidos, seja X a proporção de adultos com mais de 65 anos que estejam empregados, ou a taxa de emprego das pessoas mais velhas. Então, X estará restrita a um valor entre zero e um. Suponha que a função de distribuição cumulativa de X seja dada por $F(x) = 3x^2 - 2x^3$ para $0 \leq x \leq 1$. Encontre a probabilidade de que a taxa de emprego das pessoas mais velhas seja de pelo menos 0,6 (60%).

B.5 Um pouco antes da seleção dos jurados para o julgamento por assassinato de O. J. Simpson em 1995, uma pesquisa de opinião constatou que 20% da população adulta acreditava que Simpson era inocente (após a maioria das provas físicas do caso ter se tornada pública). Ignore o fato de que esses 20% sejam uma estimativa baseada em uma subamostra da população; a título ilustrativo, considere esse número como a porcentagem verdadeira das pessoas que achavam que Simpson era inocente, antes da seleção dos jurados. Assuma que os 12 jurados tenham sido selecionados aleatória e independentemente da população (embora isso não tenha sido verdade).

(i) Encontre a probabilidade de que no júri havia pelo menos um membro que acreditava na inocência de Simpson antes da seleção dos jurados. [*Sugestão*: Defina a variável aleatória

X Binomial(12;0,20) como o número de jurados que acreditavam na inocência de Simpson.]

(ii) Encontre a probabilidade de que no júri havia pelo menos dois membros que acreditavam na inocência de Simpson. [*Sugestão*: $P(X \geq 2) = 1 - P(X \leq 1)$, e $P(X \leq 1) = P(X = 0) + P(X = 1)$.]

B.6 (Exige cálculo integral.) Seja X a sentença prisional, em anos, das pessoas condenadas por roubo de veículos em determinado estado dos Estados Unidos. Suponha que a fdp de X seja dada por

$$f(x) = (1/9)x^2, \ 0 < x < 3.$$

Use integração para encontrar a sentença prisional esperada.

B.7 Se um jogador de basquetebol converte 74% dos lances livres que faz, então, em média, quantos ele converterá em um jogo com oito lances livres?

B.8 Suponha que um aluno de uma universidade esteja fazendo três cursos, em determinado semestre: um curso de dois créditos, um curso de três créditos e um curso de quatro créditos. A nota esperada no curso de dois créditos é 3,5, enquanto a nota esperada nos cursos de três e quatro créditos é 3,0. Qual será a nota média global esperada no semestre? (Lembre-se de que cada curso é ponderado pela sua participação no número total de unidades de créditos.)

B.9 Seja X o salário anual dos professores universitários nos Estados Unidos, medido em milhares de dólares. Suponha que a média salarial seja 52,3 com um desvio-padrão de 14,6. Encontre a média e o desvio-padrão quando o salário é medido em dólares.

B.10 Suponha que, em uma grande universidade, a nota média de graduação, *nmgrad*, e a nota do exame vestibular, *sat*, estejam relacionadas pela esperança condicional $E(nmgrad|sat) = 0{,}70 + 0{,}002\ sat$.

(i) Encontre a *nmgrad* esperada quando $sat = 800$. Encontre $E(nmgrad|sat = 1.400)$. Comente sobre a diferença.

(ii) Se a *sat* média da universidade for 1.100, qual será a *nmgrad* média? (*Sugestão*: use a propriedade EC.4.)